Solution of Simultaneous Equation Models in High Performance Systems

José-Juan López-Espín[1] and Domingo Giménez[2]

[1] Departamento de Estadística, Matemáticas e Informática
Universidad Miguel Hernández, Alicante, Spain
jlopez@uhm.es
[2] Universidad de Murcia, Departamento de Informática y Sistemas
30071 Murcia, Spain
domingo@dif.um.es

Abstract. Simultaneous Equation Models (SEM) have been used traditionally in econometrics, but recently they have begun to be used in networks simulation, biological microsystems, psychology, etc. In some cases the systems to solve are very large and it would be interesting to have efficient high performance algorithms to solve them.
In this work the solution of SEM in high performance systems is analyzed. Parallel algorithms of the Ordinary Least Squares (OLS), the Indirect Least Squares (ILS) and the Two-stage Least Squares (2SLS) methods are developed. Because least squares algorithms are repeatedly applied in the solution of SEM, the parallel algorithms must be designed allowing the reutilization of data shared by different problems. Algorithms for shared memory (using OpenMP) and distributed memory (using MPI) are studied theoretically and experimentally. The algorithms make an extensive use of basic libraries like BLAS and LAPACK to obtain efficient and portable versions.

1 Simultaneous Equation Models

1.1 Available software

Traditionally the Simultaneous Equation Models have been used in econometrics [1, 2], but recently they have begun to be used in networks simulation [3], biological microsystems, psychology [4], etc. In 1994 QMS began the development of software for this type of systems, with Eviews 1.0. The last version is 6.0. Eviews includes linear regression techniques, solution of Simultaneous Equations Models, times series, and other econometric problems [5]. One free tool for econometric is Ox [6]. This software includes fewer problems than Eviews, althougt it includes those studied in this work. Other statistics packages include 2SLS, but this is used in the solution of linear regression equations, and not for solving Simultaneous Equation Models. For example, SPSS 12.0 includes 2SLS.

All this software is for sequential processors, but in some cases the models to solve are very large and it may be preferable to solve them using high performance computers. The only previous works we know of on parallel algorithms to solve this type of problems are those by Kontoghiorghes [7], but he only studies OLS and 3SLS, which works with complete information and is not commonly used.

1.2 The problem

A Simultaneous Equations System is a regression equation system where three types of variables appear:

- Endogenous variables. They are internal variables of the system, which influence, and are influenced by, the other variables.
- Predetermined variables. They influence the system, but are not influenced by the system. They can be: exogenous (external to the system) and endogenous in the time (they are lagged endogenous variables, which influence the system, but can not be influenced because their data are earlier).
- White noise (random variables). They form the non controllable part of the regression equation.

The scheme of a system with M equations, M endogenous variables and k predetermined variables is:

$$\begin{array}{l} Y_{1,t} = \beta_{1,2}Y_{2,t} + \beta_{1,3}Y_{3,t} + \ldots + \beta_{1,M}Y_{M,t} + \gamma_{1,1}X_{1,t} + \ldots + \gamma_{1,k}X_{k,t} + u_{1,t} \\ Y_{2,t} = \beta_{2,1}Y_{1,t} + \beta_{2,3}Y_{3,t} + \ldots + \beta_{2,M}Y_{M,t} + \gamma_{2,1}X_{1,t} + \ldots + \gamma_{2,k}X_{k,t} + u_{2,t} \\ \ldots \\ Y_{M,t} = \beta_{M,1}Y_{1,t} + \ldots + \beta_{M,M-1}Y_{M-1,t} + \gamma_{M,1}X_{1,t} + \ldots + \gamma_{M,k}X_{k,t} + u_{M,t} \end{array} \tag{1}$$

where $Y_1, Y_2, \ldots, Y_M$ are endogenous variables, $X_1, X_2, \ldots, X_k$ are predetermined variables, and $u_1, u_2, \ldots, u_M$ are random variables.

Equation 1 can be represented in matrix form as:

$$BY_t + \Gamma X_t + u_t = 0 \tag{2}$$

The problem consists of obtaining $\beta_{1,1}, \beta_{1,2}, \ldots, \beta_{M,M-1}, \gamma_{1,1}, \ldots, \gamma_{M,k}$ from a representative sample of the model.

The structural model (equation 2) can be expressed in reduced form:

$$Y_t = \Pi X_t + v_t \tag{3}$$

1.3 Estimation by Indirect Least Squares

The ILS technique needs the equation to be exactly identified, which means the values of B and Γ can be univocally obtained. The technique estimates the values of Π by OLS, and from Π obtains the values of the structural form of the equation it is solving. Thus, ILS solves the equations systems $-B_i\Pi = \Gamma$, with B_i being the row of B corresponding to the equation which is being solved, and Π is common to all the equations. In the first step, matrix Π is obtained and it is used in successive steps.

1.4 Estimation by Two-stage Least Squares

With 2SLS it is not necessary for the equation to be exactly identified.

In an equation, the problem of the correlation between random and endogenous variables is avoided by substituting the original variable by a new variable called `proxy`. The variable is obtained by applying OLS to the predetermined variables in the system. Once all the endogenous variables have been substituted, OLS is applied to the equation.

2 Solution of Simultaneous Equations Models on parallel systems

Parallel versions for Shared Memory (OpenMP) and Distributed Memory (MPI) have been developed. The loops in the programs have different costs in different iterations, and some results obtained in one iteration can be used in successive iterations. This makes it difficult to obtain perfectly balanced parallel versions. Parallelization can be made at different levels: in the basic matrix operations using PBLAS, dividing the work in the loops between the threads in the system, etc. The parallelization has been made each time at the highest possible level.

2.1 Paralelization of ILS

The solution of only one equation by ILS has been parallelized. The parallelization of the first iteration has been entrusted to the basic functions (matrix multiplication, inverse, etc). The works in successive iterations are distributed between the different threads. The same idea is applied to solve a complete system. When ILS is applied to solve all the equations in the system the computation of matrix Π is made only once.

2.2 Paralelization of 2SLS

The solution of only one equation by 2SLS has been parallelized. 2SLS is divided in two parts: computation of the `proxys`, and the application of a special OLS routine. In the loop to calculate the `proxys` not all the iterations have the same cost, and this must be borne in mind to obtain an efficient parallel algorithm. The difference in the costs of the computations is caused by the fact that some results of OLS are used in successive iterations. Using the same idea as in ILS the first iteration is parallelized at a low level (matrix multiplications), obtaining a matrix which will be used in the successive iterations. Once the first iteration has been performed, the other iterations are divided between the available threads. When the loop has been made, the parallelization of the special OLS is made at a low level.

2.3 Parallelization of a Simultaneous Equations System

To solve a complete Simultaneous Equations System it is better to parallelize at a higher level. In a system there can appear non identified, overidentified and exactly identified equations. For non identified equations, no estimation can be made, for overidentified equations it is necessary to apply 2SLS, and for exactly identified equations ILS or 2SLS can be applied. In a large system, equations of all the types will appear, and ILS and 2SLS are applied. In 2SLS, each thread works in the solution of the equations without sharing information with the other threads. In ILS, matrix Π is common to all the equations. The first iteration is solved as described, then the loop to solve the other equations is solved in parallel, and all the threads use the matrix computed.

In distributed memory the data are initially in processor P_0, which distributes the data between all the processors in the system. After receiving the data, each processor solves the equations which it has assigned. Some equations are solved by ILS and others by 2SLS (the rest of the equations are non identified).

2.4 Experimental results

Experiments have been performed in a shared memory system and a cluster. The results obtained in both systems with ILS, 2SLS and with a complete SEM are satisfactory. The table shows the execution time and the speed-up obtained when varying the number of processors and the number of endogenous variables, with ILS in the cluster.

size:	500		1000		2000	
proc.	time	Sp	time	Sp	time	Sp
1	18.16		346.77		5320.91	
3	6.64	2.73	117.20	2.96	1823.38	2.92
5	4.28	4.24	71.32	4.86	1065.92	4.99
10	3.74	4.86	36.84	9.41	559.36	9.51
18	2.23	8.14	24.55	14.13	310.05	17.16

Acknowledgment

This work has been funded in part by the Spanish MCYT and FEDER under Grants TIC2003-08238-C02-02.

References

1. Greene, W.: Econometric Analysis, third edition, Prentice Hall, 1998.
2. Gujarati, D.: Basic Econometrics, McGraw Hill, 1995.
3. Harchol-Balter, M., Black, P. E.: Queueing Analysis of Oblivious Packet-Routing Networks, SODA: ACM-SIAM Symposium on Discrete Algorithms, 1993.
4. http://www.sswr.org/papers2002/616.htm
5. http://www.eviews.com/
6. http://www.doornik.com/
7. Kontoghiorghes, E. J.: Parallel Algorithms for Linear Models: Numerical Methods and Estimation Problems, Kluwer Academic Publishers, 2000.