Obtaining simultaneous equation models from a set of variables through genetic algorithms

José J. López
Universidad Miguel Hernández (Elche, Spain)
Domingo Giménez
Universidad de Murcia (Murcia, Spain)

ICCS 2010

Contents

Introduction

- Simultaneous Equation Models (SEM) have been used in econometrics for years. Nowadays they are used in medicine, network simulation, and even in the study of the divorce rate.
- Traditionally, SEM have been developed by people with a wealth of experience in the particular problem represented by the model.
- The objective is to develop an algorithm which, given the endogenous and exogenous variables, finds a satisfactory SEM.
- The space of the possible solutions is very large and exhaustive search methods are not suitable here.
- A combination between a genetic algorithm and a greedy method is studied.

Simultaneous Equations Models

The scheme of a system with *N* equations, *N* endogenous variables, *K* exogenous variables and *d* sample size is (*structural* form)

$$Y_1 = \beta_{12}Y_2 + \beta_{13}Y_3 + ... + \beta_{1N}Y_N + \gamma_{11}X_1 + ... + \gamma_{1K}X_K + u_1$$
$$Y_2 = \beta_{21}Y_1 + \beta_{23}Y_3 + ... + \beta_{2N}Y_N + \gamma_{21}X_1 + ... + \gamma_{2K}X_K + u_2$$
$$\dots$$
$$Y_N = \beta_{N1}Y_1 + \beta_{N2}Y_2 + ... + \beta_{NN-1}Y_{N-1} + \gamma_{N1}X_1 + ... + \gamma_{NK}X_K + u_N$$

where Y_i, X_j and u_i are *dx1* $i = 1...N,\ j = 1...K$

These equations can be represented in matrix form

$$BY + \Gamma X + u = 0$$

The problem: Find the best SEM given a set of values of variables

- One model is considered better than another if it has a lower criteria parameter.
- AIC and BIC are two of the most used methods for comparing models.

$$AIC = d \ln |\hat{\Sigma}_e| + 2\sum_{i=1}^{N}(n_i + k_i - 1) + N(N+1)$$

$$BIC = d \ln |\hat{\Sigma}_e| + (\ln d)\left(\sum_{i=1}^{N}(n_i + k_i - 1) + 0.5N(N+1)\right)$$

  - d is the sample size, n_i and k_i the number of endogenous and exogenous variables in equation i, and $\hat{\Sigma}_e$ is the covariance matrix of the errors.

Genetic Algorithms for selecting the best SEM

- Each chromosome represents one candidate.
- A chromosome is defined as a matrix with *N* rows and *N+K* columns.
- In each row, an equation is represented using ones and zeros.
- If variable *j* appears in equation *i*, the value for the *(i,j)* position in the chromosome is one, and zero if not.
- The first *N* columns of a chromosome represent the endogenous variables and the other *K* columns represent the exogenous ones.

For example, in a problem with *N=2* endogenous variables (Y_1 and Y_2) and *K=3* predetermined variables (X_1, X_2 and X_3):

$$y_1 = \beta_{1,2} y_2 + \gamma_{1,1} x_1 + \gamma_{1,2} x_2 + u_1$$

$$y_2 = \beta_{2,1} y_1 + \gamma_{2,3} x_3 + u_2$$

$\longrightarrow$

11110

11001

Defining a valid chromosome

- The model has to have at least one equation.
- If the *(i,i)* element is zero, the column *i* will have only zeros.
- Each equation in the model must have at least two variables.
- Rank condition: Equation *i* is identified if it is possible to find a *(N-1) x(N-1)* matrix with full range where the columns are the unknown variables

$$\gamma_{1,1},\ldots,\gamma_{N,K},\beta_{1,2},\ldots,\beta_{N,N-1}$$

that do not appear in the equation.

- The number of comparisons when evaluating a chromosome is :

$$N^2 + N(N+K) + N\frac{(K+N-2)!}{(K-1)!}$$

Evaluating a chromosome

- The algorithm on the right shows the scheme of the fitness function of a chromosome.

- The cost of evaluating a chromosome is :

$$\approx \mathrm{O}(K^2 Nd + K^3 N)$$

1. **BUILD** the system using chromosome c and the set of variables Y and X
2. **SOLVE** the system
3. **COMPUTE** the error between the variables Y and its estimation
4. **COMPUTE** AIC or BIC

Initialization and EndConditions

- Each chromosome is generated according to the algorithm on the right.
- The population size (called *PopSize*) is stated at the beginning.
- The process is repeated until it reaches a maximum number of iterations, called *MaxIter*, or the best fitness is repeated over a number of successive iterations, called *MaxBest*.
- Both parameters are stated at the beginning of the algorithm.

1. **GENERATE** the $N(N+K)$ elements randomly (with the same probability of zeros and ones)

{C1 AND C2 CONDITIONS}

2. **IF** N or $N-1$ elements $e_{(i,i)}$ are zero with i=1,…,N
3. invert all the elements $e_{(i,i)}$ with i=1,…,N
4. **END IF**

{C3 CONDITION}

5. **FOR** i=1…N
6. **IF** the element $e_{(i,i)}$ is zero
7. make all the elements zero in column i
8. **END IF**
9. **END FOR**

{C4 CONDITION}

10. **FOR** i=1…N
11. **IF** equation i fails the range condition
12. generate randomly this equation (row i) and go to 2
13. **END IF**
14. **END FOR**

Crossover

- Three sorts of crossover are studied:
  - Single Point (SP)
  - Single Point considering equations (SPCE)
  - Inside an Equation (IE)

parents		**SP** e = 10		**SPCE** e = 1		**IE** e = 2, v1 = 2, v2 = 3	
parent1	parent2	child1	child2	child1	child2	child1	child2
11110110	10100100	11110110	10100110	11110110	10100100	11110110	10100100
11110101	01110100	11110100	01110101	01110100	11110101	11110100	01110101
01110110	11110110	11110110	01110110	11110110	01110110	01110110	11110110

problem size			crossover **SP**			crossover **SPCE**			crossover **IE**		
N	***K***	***d***	t	iter	*FF*	t	iter	*FF*	t	iter	*FF*
10	15	50	3.03	48	**2683.13**	5.11	97	2732.90	**0.66**	20	2833.41
15	20	50	8.00	62	4548.68	6.73	53	**4540.93**	**1.94**	40	4709.50
30	40	100	58.33	50	**21937.02**	87.54	72	22120.10	**9.47**	17	22765.68
40	50	100	325.87	111	**30956.78**	294.19	102	31262.20	**64.41**	24	32975.04

Mutation

- A small probability of mutation is considered in each iteration.
- A chromosome of the new subset generated in the crossover is chosen randomly, and an equation and a variable are generated randomly. Then, the element is inverted.
- PROBLEM: When a chromosome is mutated and then situated in a different part of the set of solutions, it does not normally have enough quality to survive to create new chromosomes in this area, and perhaps a better solution is close to it.

Greedy Method

To avoid this problem, a greedy method is used in the mutation, following the algorithm on the right.

- A chromosome *c* is chosen randomly from the population
- An equation *e* and a variable *v* are chosen randomly in *c* and the element (*e,v*) is inverted obtaining c_1.
- The best chromosome in the neighbourhood (those obtained by inverting only one element) is search.
- If the best chromosome coincides with c_1, the loop ends and it is included in the population.
- If not, c_1 is substituted by the best chromosome found and the process continues.
- This process is repeated until NEG (Number of Equations in Greedy) different equations are generated.

		N=10, K=20		N=20, K=30	
Mode	**NEG**	**AIC**	**time**	**AIC**	**time**
without greedy method	-	2138.93	5.10	4658.06	15.41
with greedy method	1	2143.54	9.79	4710.53	49.14
	[N/2]	1491.13	12.62	3072.98	102.23
	[3N/4]	-680.61	27.48	811.65	227.35
	N	-3586.46	34.17	-4920.01	449.78

Experimental Results

- The error shown is the sum of the squares of the differences between the values observed of the main endogenous variables and those obtained by the estimation of these, divided by the values observed.

- In most cases BIC obtains models with lower error than AIC.

- But the behaviour of BIC is irregular because in some cases models with lower BIC and higher error are obtained.

			PopSize=100		PopSize=500	
N	K	Sigma	Error_AIC	Error_BIC	Error_AIC	Error_BIC
30	40	0	$1.47_{0.72}$	$\mathbf{1.24_{0.65}}$	$\mathbf{1.31_{0.31}}$	$1.33_{0.54}$
30	40	0.01	$1.17_{0.32}$	$\mathbf{0.99_{0.39}}$	$0.88_{0.28}$	$\mathbf{0.87_{0.36}}$
30	40	0.1	$1.06_{0.32}$	$\mathbf{0.92_{0.42}}$	$\mathbf{0.91_{0.35}}$	$0.95_{0.31}$
40	50	0	$2.29_{0.52}$	$\mathbf{2.01_{0.43}}$	$2.29_{0.64}$	$\mathbf{2.28_{0.78}}$
40	50	0.01	$1.64_{0.46}$	$\mathbf{1.58_{0.40}}$	$\mathbf{1.59_{0.49}}$	$1.62_{0.27}$
40	50	0.1	$1.64_{0.37}$	$\mathbf{1.54_{0.34}}$	$1.56_{0.38}$	$\mathbf{1.31_{0.19}}$

Experimental Results

- The costliest parts of the genetic algorithm are *Evaluate*, *Crossover* and *Mutate*, and have been paralleled simply by assigning some chromosomes to each processor.
- The algorithm is stopped when the maximum number of iterations (*MaxIter*) is reached.

				1proc	2proc		4proc		8proc	
PopSize	N	K	d	time	time	sp	time	sp	time	sp
100	10	20	100	17.25	10.61	1.63	6.48	2.66	3.73	4.63
100	20	30	100	123.04	63.74	1.93	33.41	3.68	20.72	5.94
100	30	40	100	717.75	370.99	1.94	190.42	3.77	98.48	7.30
500	10	20	100	71.2	41.74	1.71	24.66	2.89	16.29	4.37
500	20	30	100	280.09	144.82	1.93	97.48	2.87	54.06	5.18
500	30	40	100	1309.45	682.78	1.92	344.18	3.81	180.86	7.24

Conclusions and Future works

Conclusions

- An algorithm to obtain a satisfactory Simultaneous Equation Model from a set of variables is studied.
- Genetic and greedy method are combined to avoid to fall into local minima and to speed up the convergence.
- A shared memory version, which allows us to efficiently use multicore processors in the solution of the problem, has been developed.

Future Works

- Application to real problems.
- Develop a hybrid (message-passing plus shared memory) algorithm.
- Use and compare other criteria parameters.
- Use Other metaheuristic methods (Scather Search, GRASP,...)

References

- Akaike, H., Information theory and an extension of the maximum likelihood principle. In: B.N. Petrov, Csaki F. (Ed.), Proc. 2nd Int. Symp. on Information Theory, Akademiai Kiado, Budapest, 267 281, 1973.
- Bedrick, E.J., Tsai, C.-L. Model selection for multivariate regression in small samples. Biometrics, 50, 226-231, 1994.
- Bozdogan, H., Haughton, D. Informational complexity criteria for regression models. Computational Statistics and Data Analysis, 28, 51-76,1998.
- Fujikoshi, Y., Satoh, K. Modified AIC and Cp in multivariate linear regression. Biometrika, 84 (3), 707-716, 1997.
- Gorobets,A., The Optimal Prediction Simultaneous Equations Selection, Economics Bulletin, 36(3), 1-8, 2005.
- Gujarati, D. 1995. Basic Econometrics, McGraw Hill.
- Mitchell, M. 1998. An Introduction to Genetic Algorithm, MIT Press.
- Shi, P., Tsai, C.-L., . A note on the unification of the Akaike information criterion. J.R. Statist. Soc. B, 60 (3), 551-558, 1998.

Experimental Results

- Experimental results have been obtained in an Intel Itanium 2 system equipped with four dual-core 1.4 GHz Montecito processors.
- To analyze the goodness of the solutions and to compare AIC and BIC criteria
  - A valid chromosome is generated randomly (this chromosome represents the real SEM to be obtained).
  - The exogenous variables are generated randomly and the endogenous variables are calculated using equation

$$Y = \Pi X + v$$

$$(BY + \Gamma X + u = 0, \Pi = -B^{-1}\Gamma, v = -B^{-1}u)$$

with $v \sim N(0, \sigma)$