MODEL SELECTION VIA GENETIC ALGORITHMS

Eduardo Acosta-González
Departamento de Métodos Cuantitivos en Economía y Gestión
Universidad de Las Palmas de Gran Canaria
correo-e: eacosta@dmc.ulpgc.es

Fernando Fernández-Rodríguez
Departamento de Métodos Cuantitivos en Economía y Gestión
Universidad de Las Palmas de Gran Canaria
correo-e: ffernandez@dmc.ulpgc.es

RESUMEN

We provide a new simple procedure for selecting econometric models. It is based on a heuristic approach called genetic algorithms which are used to explore the universe of models made available by a general unrestricted model. This search process of the correct model is only guided by the Schwarz information criterion, which acts as the lost function of the genetic algorithm in order to rank the models. Our procedure shows good performance relative to other alternative methodologies.

JEL classification: C20; C61; C63
Keywords: Genetic algorithms, Data mining, Regressors selection, Model selection.

[*] This research was supported by the Spanish Ministry of Science and Technology through the project BEC2001-3777.

1. INTRODUCTION

Data mining consists of the extraction of hidden predictive information from large databases through several partially automated research strategies. Initially, typical data mining activities consisted of selecting from alternative models with Stepwise regression programs based on significant t-coefficients and finding high R^2, but it turns out that these procedures are not very useful as powerful selection criteria. Exclusive reliance on R^2 and t-statistics as model choice criteria can easily lead to the selection of poor models.

A seminal work in this area was published by Lovell (1983) which sought to evaluate the ability of some model selection methods to locate a single conditional equation from a large macroeconomic database, containing up to 40 regressors, including lags. Lovell found that the reduction in the cost of several procedures of modeling per se through data mining techniques, are bad because the modeling process has to be matched by a proportional increase in our knowledge of how the economy actually works. Specifically, Lovell (1983) has shown that the data mining activities of a researcher erode the levels of significance in hypothesis tests when the researcher chooses to consider the best result in isolation.

More recently the London School of Economics (LSE) approach presented a variety of competing econometric methodologies known as the general to specific modeling approach that amounts to systematized data mining. The art of model specification in the LSE framework is to seek out models that are valid parsimonious restrictions of the complete general model, and that are not redundant in the sense of having an even more parsimonious model nested within them that is also a valid restriction of the complete general model. So, the LSE approach permitted us to reconsider the model construction, emerging literature favorable to data mining activities. A seminal paper in that sense was that of Hoover and Perez (1999) (HP). HP developed a mechanical algorithm which mimics some aspects of the search procedures used by LSE practitioners, simulating general to specific selection for linear dynamic regression models, releasing the model selection strategy on residual diagnostic and hypothesis testing on coefficients. In the same direction Hendry and Krolzig (1999) suggested several alternative developments of Hoover and Perez algorithm, improving their work and supplying a computer automation of general to specific model selection procedure, which is commercially available in the software package PcGets. Additional improvements were introduced by Hendry and Krolzig (2001).

One important part of all previous papers on computer automation model selection is the search path for selecting the final model avoiding the search in all submodels which is prohibitively expensive when the number of potential regressors is large. The selection of variables procedure used on previous papers are based principally on the t-statistic as in the Stepwise and backward elimination procedures, but in each step many diagnostic test are used to check models. This improves considerably the results.

There are other researches which don't center the model selection strategy on residual diagnostic and hypothesis testing on coefficients. That is the case with Hansen (1999) who provides a procedure using only on a simple information criterion. Also, a very recent paper by Perez-Amaral et al. (2003) uses out of sample performance measures for model selection. In this research we provide a new model selection procedure based on Genetic Algorithms guided by the Schwarz information criterion. It shows higher capability in model selection than in other methods known at the moment, performing as well as, if not better, than other alternative methodologies.

The remainder of this paper has been organized as follows. In Section 2 the problem of selecting regressors is introduced. In Section 3 a brief review of genetic algorithms is presented, stressing their applications to the problem of selecting regressors. Section 4 focuses on the empirical results showing that GASIC improves on previous procedures. Section 5 presents the conclusions.

2. THE PROBLEM OF SELECTING REGRESSORS

The contribution of our paper to model building is a powerful procedure of selecting regressors which permits a very good model selection performance using a simple information criterion. In building a multiple regression model, a crucial problem is the selection of regressors to be included. If a lower amount of regressors are selected in the model, the estimate of the parameters will not be consistent and if a higher amount is selected, its variance will increase.

Given a dependent variable Y and a set of potential regressors $X_1,....,X_K$, the problem is to find the best submodel of the form:

$$Y = \beta_0 + \beta_1 X_{i_1} + ... + \beta_K X_{i_K} + \varepsilon \text{ , where } \{i_1, i_2 ...., i_K\} \subseteq \{1, 2, ...K\}$$

There are 2^K possible submodels, a wide variety of selection procedures of all possible submodels have been proposed including Akaike Information Criterion [Akaike. (1973).], Mallows criterion [Mallows (1973)], Schwarz Information Criterion [Schwarz (1978)] or Hannan-Quinn Information Criterion [Hannan and Quinn, (1979)]. When K is high the computational requirements for these procedures can be prohibitive because the number of models becomes infeasible. For example, in the context of the Lovell research the number of possible models is 2^{40}.

In order to resolve this intractable problem, several heuristic methods addressed to restrict attention to a smaller number of potential subset of regressors are usually employed by practitioners. Such heuristic procedures, rather than search through all possible models, seek a good path through them. Some of the most popular are the Stepwise procedures, such as forward selection or backward elimination, sequentially include or exclude variables based on t-ratio statistic considerations [see Miller (2002) as a review of subset selection in regression].

The Stepwise regression procedure starts off by choosing a model containing the single best regressor and then attempts to build up with subsequent additions of regressors one at a time as long as these additions are worthwhile. The order of addition is determined by using the t-statistics values to select which variable should enter next. After a variable has been added, the model is examined to see if any regressor should be deleted. The procedure terminates when all regressors not in the model are insignificant at a chosen significant level. Another procedure is the backward elimination. In this case, it starts off by estimating a model with all the potential regressors. Then the regressor having the smallest t statistics is removed from the insignificant regressors at a chosen significant level. The procedure continues until no regressors remaining in the model can be removed.

Perhaps the more well known heuristic procedure used in Econometrics is the path search considered in Hoover and Perez (1999) which is devoted to avoiding the search in all submodels which is prohibitively expensive because it requires computing 2^{40} distinct submodels. The essential characteristic of the Hoover and Perez algorithm and its derivatives is the choice of a battery of tests (residual diagnostics and hypothesis testing on coefficients), a measure of fit and a search path. So, the HP algorithm examines all models along the search path, selecting the best-fitting model among those models which are not rejected by the tests. The path search is an essential part of the HP algorithm. The HP search path is as follows: The variables of the general specification are ranked in ascending order according to their t-statistics. For each replication, 10 search paths are examined. Each search path begins with the elimination of one of the variables in the subset with the 10 lowest (insignificant) t-statistics. The first search begins by eliminating the variable with the lowest t-statistic and re-estimating the regression. This re-estimating regression becomes the current specification. The search continues until it reaches a terminal specification. Multiple paths can lead to multiple models, so after all paths are explored; Hoover and Perez select the model that fits best.

The HP search path has been improved by Hendry and Krolzing (1999, 2001), and Krolzig and Hendry (2001) proposing potential improvements that include trying all single-variable deletion starting points, and feasible block deletion. So, groups of variables are tested in the order of their absolute t-values, commencing with a block where all the p-values exceed 0.9, and continuing down towards the pre-assigned selection criterion, when deletion must become inadmissible. Besides, in addition to individual coefficients, blocks of variables constitute feasible paths. These additional tests operate like the block F-test along search paths. Then all paths that also commence with an insignificance t-deletion are explored. All of it increases the number of paths searched. If at the end of this procedure there is more than one model they are tested against their union, and finally, if there is a set of non- dominated terminal models they are selected using information criteria instead of using the standard error of regression as does HP algorithm. The descriptive of the Hendry and Krolzig basic algorithm and some recent changes can be found in Hendry and Krolzig (2003).

Hansen (1999) suggested simplifying the HP model selection procedure, although proposing a search path too simple and not always suitable. Based on numerical evidence Hansen claims that simple and elegant Schwarz Information Criterion (SIC) work at least as well, if no better, than the complicated algorithm attributed to the LSE methodology

Although the SIC looks like one of the most promising criterion for selecting models, Hansen's reduction of the maximum number of regressors to $K'=10$ is arbitrary and it can't ever be justified in advance. The contribution of our research on model selection is to provide a new search path algorithm avoiding the main inconvenience when the number K of potential regressors is large, proposing a heuristic approach called Genetic Algorithm (GA henceforth). This search process is only guided by the Schwarz Information Criterion (SIC), which acts as the lost function of the GA. So, we will designate the acronym GASIC to our selecting model procedure. GASIC permits exhaustive exploration of the most promising models of the search space without a cumbersome time consuming process and shows higher capability in model selection than other methods known at the moment, performing as well as, if not better, than complicated search algorithms.

There exists a previous paper by Beenstock and Szpiro (2002) who proposed the use of GA to estimate dynamical nonlinear time series models from non stationary data. Our paper is devoted to the specific problem of variables selection in linear models and there are substantial differences with Beenstock and Szpiro (2002) at least in two aspects. On the one hand, Beenstock and Szpiro (2002) look for very general unrestricted functional forms. On the contrary, our research only selects linear models starting off a correct general specification like the approach of London School of Economics. So, Beenstock and Szpiro (2002) operate on strings representing functional forms using a variation of GA developed by Koza (1992) called Genetic Programming. In contrast, we use the binary version of GA suggested originally by Holland (1975), with the end of selecting explanatory variables, providing a regressors selector algorithm by linear models. Therefore the Beenstock and Szpiro (2002) procedure is more general because in addition to searching over lags and variables, they search for functional forms. In contrast, GASIC is devoted to the problem of selecting linear models although it can be extended to the framework of semi parametric models. GASIC is shown as a more parsimonious and robust tool able to obtain predictions as much as undertaking a simple structural analysis. It is the fixed and closed linear form of our model which makes possible the structural analysis.

On the other hand, the Beenstock and Szpiro (2002) algorithm does not generally converge towards a unique true model and the result generally depends upon the reseeding of the initial population. However, the closed linear functional form in GASIC shows higher robustness to reseeding in the final solution, as may be seen in our simulations.

3. SELECTING REGRESSORS VIA GENETIC ALGORITHMS

The new approach for selecting regressors proposed in this paper is based on an heuristic optimization procedure called GA. Before moving on to explain our method, let's summarize the main aspects of GAs. GA are a class of optimization technique, based on principles of natural evolution developed by Holland (1975) which try to overcome problems of traditional optimization algorithms, like no continuity or differentiability of the loss function.

A GA starts with a population of randomly generated solution candidates, which apply the principle of fitness to produce better approximations to optimal solution. Promising solutions as represented by relatively better performing solutions, are selected and breeding them together through a process of binary recombination referred to as crossover inspired by Mendel's natural genetics. The objective of this process is to generate successive populations solutions that are better fitted to the optimization problem than the solutions from which they were created. Finally, random mutations are introduced in order to avoid local optima.

GA have been applied to a variety of problems in a diverse range of fields; they are most effectively used in situations where the space of possible solutions to an optimization problem is too large to be handled efficiently by standard procedures, or when it is in some sense badly behaved e.g. non-differentiable, or possessing multiple local extrema [See Goldberg (1989) as a general reference].

Although the use of GA has been generalized on a variety of problems in a diverse range of fields, applications of GA in econometrics are scarce, in this sense the work of Dorsey and Mayer (1995) stands out where several optimization problems in econometric estimation were carried out.

Our model selection procedure for real data may be resumed as follows:

a) Check the congruence of the general unrestricted model using a battery of mis-specification tests as Hendry and Krolzig (2001) point out. Empirically, the general unrestricted model would be revised if such tests were rejected. In our results we omitted this step because we work with simulated data bases and our contribution refers to the following step b.

b) Select a submodel of the general unrestricted model using GASIC where the lost function which ranks models is the Schwarz Information Criterion (SIC).

The basic steps in constructing GASIC are the followings:

Step 1. *Initial Population:* Generating a population of random solutions to the optimization problem usually called individuals. These solution candidates, also called chromosomes, are usually represented by vectors, all of the same length, consisting of binary digits. When the parameters are naturally discrete the binary GA fits nicely. In K-dimensional optimization problem, a chromosome is written as an array with 1xK elements so that

$$chromosome = \left[p_1, p_2, ..., p_K \right]$$

where p_i , $i = 1....K$ are a binary variable taking values zero or one, and K is the number of regressors. For instance, if K=5 and the complete set of regressors (general model) are $\{ X_1, X_2, X_3, X_4, X_5 \}$ the chromosome (1,0,1,0,1) means that the subset of regressors considered is $\{ X_1, X_3, X_5 \}$. Therefore, our algorithm begins with the random selection of an initial population of binary chromosomes which represents random approaches of the General Unrestricted Model. These chromosomes act as seeds of the process. Nevertheless, the algorithm is robust in reseeding as we will show in the empirical results. In the empirical results developed in section 4, the number of chromosomes of the initial population will be 200. With the increase of this number of chromosomes, results are only improved marginally.

Step 2. *Ranking:* For every chromosome, the lost function is calculated. According to Hansen (1999) and Campos et al. (2003), we have considered a lost function provided by the Schwarz Information Criteria (SIC),

$$SIC(m) = \log \hat{\sigma}^2(m) + c \frac{\log(T) k(m)}{T} \text{ , } c = 2 \tag{1}$$

where k is the number of “ones” in each chromosome m, which represent the selected regressors, and T is the sample size. So, one solution is considered better fitted than another if the value of SIC of the first is lower.

The model selection based on minimizing the SIC is consistent, having the advantage that incorrect models are never selected asymptotically when the sample size diverges away from K. Also the correcting factor $c = 2$ avoids the possibility of over-parameterized models being asymptotically selected with positive probability, as Hansen (1999) pointed out.

The profitability of using SIC on selection models is discussed in Campos et al. (2003) who find that the significant levels embedded in the PcGets algorithm coincide in very large samples with those implicit in the SIC. A given value for c determines the implicit significance level for the SIC as a function of T and K. So, Choosing $c > 1$ is tantamount to choosing a more stringent p-value.

Step 3. *Natural Selection:* In order to simulate the process of Darwin's natural selection, chromosomes are ordered on the basis of their lost function and the worst half of individuals are discarded (deleted).The subpopulation that we save is usually called the mating pool. So, in our problem we delete those models with high SIC-statistics.

Step 4. *Pairing:* Couples of chromosomes are selected from the mating pool in order to produce two new offspring solutions. Pairing chromosomes in a GA can be carried out in a variety of methods. In our case we have carried out a random pairing process which assigns equal probability to every chromosome.

Step 5. *Mating:* This operation creates new offspring from the set of regressor´s subset selected in the pairing process. This process is called genetic recombination or crossover. It is frequent to apply the commonly used single point crossover which consists in randomly pairing chromosomes surviving the selection process, and randomly, selecting a break point at a particular position in the binary representation of each chromosome. This break point is used to separate each vector into two subvectors. The two subvectors to the right of the break point are exchanged between the two vectors, yielding two new chromosomes. For instance, let's consider a couple of chromosomes called mother and father: Mother=(0,1,0|,1,0) , Father =(**1,0,1,|0,1**), If the break point is selected after the third position in every chromosome, two new chromosomes are created through the parents: Offspring$_1$=(0,1,0|,**0,1**) and Offspring$_2$=(**1,0,1**,|1,0). Every one inheriting part of the parents` genetic material, which means that if we recombine the subset of regressors $\{X_2, X_4\}$ and $\{ X_1, X_3, X_5 \}$, we will obtain the offspring $\{X_2, X_5\}$and $\{X_1, X_3, X_4\}$.

Step 6. *Mutations:* Mutation is the process of randomly changing in the string of binary elements in a chromosome. Mutations prevents the GA from converging too quickly on a local minimum of the lost function. If the algorithm is trapped in a local optimum, the mutation randomly shifts the solution. So, a mutation occurs by randomly selecting a particular element in a particular vector. If the element is a "one" it is mutated to "zero", and *vice versa*. This occurs with a very low probability in order not to destroy promising areas of search space. In the empirical results the rate of mutation is 0.5%.

Step 7. *Convergence:* Come back to Step 2 and repeat sequentially this process getting successive generations of solutions until some convergence criterion is satisfied. The stopping criterion is usually satisfied if either the population converges to a unique solution or a maximum number of predefined generations are reached. In this paper we selected an intermediate stopping criterion consisting of stopping when the ten best behaved solutions of a generation in the GA are repeated. Additionally, in order to avoid worsening the solution, except in the first sequence of the algorithm, we proceed in step 3 as follows: In the case where the best behaved chromosome in (i)-th generation isn't as well as the best behaved chromosome in the (i-1)-th generation, this last generation will come to form the mating pool (i+1)-th generation of the GA.

4. EMPIRICAL RESULTS

In order to investigate the relative performance of GASIC we provide two different scenarios of data with different sample size: The Hoover and Perez (1999) macroeconomic data base and several of the data generating processes in the experiments simulated by Perez-Amaral et al. (2003). All of our calculations have been provided using MATLAB code.

4.1 Empirical results on Hoover and Perez (1999) data base.

We used Lovell's database modified by Hoover and Perez (1999) consisting of macroeconomics variables covering various measures of real activity, government fiscal flows, monetary aggregates, financial markets yields, labor market conditions and a time trend. We also considered the eleven specifications constructed by Hoover and Perez (1999) (Table 3). The HP search algorithm is composed of a battery of seven well-known residual diagnostics and hypothesis testing on coefficients and by a particular search path based on Stepwise consideration, which acts as an elimination procedure of variables in the general specification.

Although Hoover and Perez (1999) support the view that the particular search path mechanism plays a relatively small role in determining the search algorithm, we will see that our search procedure affects considerably the results. To assess the general to specific approach we construct a specification search for 1000 replications of the eleven specifications listed by HP. We have also distinguished five degrees of success of our algorithm in order to focus on the question of whether or not the true specification is nested within the final specification. The five Hoover and Perez (1999) categories that we reproduce are: Category 1 (*Final=True*) where the true specification is chosen. Category 2 $(True \subset Final, SER_F < SER_T)$[1], where the true specification is nested in the final specification and the final specification has the lower standard error regression. Category 3 $(True \subset Final, SER_F > SER_T)$, where the true specification is nested in the final specification and the true specification has the lower standard error regression. Category 4 $(True \not\subset Final, SER_F < SER_T)$, where an incorrect specification is chosen, the true specification is not nested in the final specification and the final specification has a lower standard error of regression than the true specification. Category 5 $(True \not\subset Final, SER_F > SER_T)$, where an incorrect specification is chosen, the true specification is not nested in the final specification, and the true specification has a lower standard error of regression that the final specification.

In Table 1 we present the results of specification searches for 1000 replications[2] of eleven specifications obtained by using GASIC. Also, all calculations carried out with SIC as a measure of goodness of fit in the GA have been repeated using the F statistics as a measure of goodness of fit. These results are not presented in this paper because they are inferior to those obtained using the SIC. In order to compare, we also present the HP and Stepwise results at 1% nominal size in Table 2 and Table 3, respectively. We decided to include the Stepwise results in order to enhance how our search path algorithm based on GA, dramatically improves it.

Tables 1 and 2 permit the comparison between GASIC with HP, for a nominal size of 1% (their best results). Looking at these tables we conclude that:

[1] SER_F refers to the standard error regression for the final specification and SER_T refers to that for the true specification.

[2] These 1000 replications refer to 1000 reseedings of the initial population of chromosomes in the GA.

Table 1. Specification searches using GASIC. (GA with lost function: SIC with c=2)

True Model	1	2	3	4	5	6	6A	6B	7	8	9	Means
Category 1	93,3	93.0	78.9	91.7	93.3	0.1	93.4	93.8	92.7	93.3	0.0	68.85
Category 2	6.7	7.0	5.6	8.3	6.7	0.0	5.7	6.1	7.2	6.7	0.0	5.45
Category 3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Category 4	0.0	0.0	3.0	0.0	0.0	6.4	0.6	0.1	0.1	0.0	7.7	1.63
Category 5	0.0	0.0	12.5	0.0	0.0	93.5	0.3	0.0	0.0	0.0	92.3	18.05
True variables	0	1	2	1	1	2	2	2	3	3	5	
Average rate of selected variables	0.08	1.08	1.95	1.11	1.08	1.07	2.07	2.07	3.08	3.07	3.12	
Average rate of true variables	---------	1.00	1.84	1.00	1.00	1.00	1.99	2.00	3.00	3.00	3.02	
Average rate of insignificant variables	0.01	0.01	0.01	0.01	0.00	0.00	0.01	0.01	0.00	0.01	0.01	0.01
Average rate of falsely significant variables	0.07	0.08	0.10	0.09	0.08	0.07	0.07	0.07	0.08	0.07	0.08	0.08
Type I error (True Size)	0.17%	0.19%	0.26%	0.24%	0.19%	0.18%	0.19%	0.17%	0.22%	0.18%	0.24%	0.20
Power	--------	100.0%	92.2%	100.0%	100.0%	50.1%	99.6%	100.0%	100.0%	100.0%	60.4%	90.23
Frequency true variables included (percent)												
True Model	1	2	3	4	5	6	6A	6B	7	8	9	
True variables number												
3					100	0.1	99.1	99.9		100	0.0	
11				100		100	100	100	100		100	
21										100	1.2	
29									99.9		99.9	
37		100	94.8						100	100	100	
38			89.6									

Hoover and Perez (1999) used 40 dependent variables as candidates to take the part of the real model, and these variables are correlatively numbered. The number of the variables appearing in the table are the true variables used in the models´ generation.

As in Hoover and Perez (1999) we define: Size=falsely significant variables /(total candidates-possible true variables) and Power = 1-(possible true variables-true variables selected)/possible true variables.

Table 2. Specification searches using Hoover Perez Algorithm (at 1% nominal size)

True Model	1	2	3	4	5	6	7	8	9	Means
Category 1	79.9	0.8	70.2	80.2	79.7	0.7	24.6	78.0	0.8	46.1
Category 2	20.1	99.2	19.0	19.6	20.2	0.1	57.4	21.7	1.3	28.7
Category 3	0.0	0.0	0.2	0.1	0.1	0.0	0.0	0.2	0.6	0.1
Category 4	0.0	0.0	3.7	0.1	0.0	56.3	13.0	0.1	77.0	16.7
Category 5	0.0	0.0	6.9	0.0	0.0	42.9	5.0	0.0	20.3	8.3
True variables	0	1	2	1	1	2	3	3	5	
Average rate of true variables		1.00	1.89	0.99	1.00	1.01	2.82	3.00	2.86	
Average rate of insignificant variables	0.01	0.07	0.04	0.05	0.04	0.02	0.11	0.05	0.06	0.05
Average rate of falsely significant variables	0.28	2.24	0.35	0.29	0.28	0.24	1.12	0.33	1.14	0.70
Type I error (True Size)	0.7%	5.7%	0.9%	0.8%	0.7%	0.6%	3.0%	0.9%	3.2%	1.8%
Power		100.0%	94.7%	99.9%	100.0%	50.3%	94.0%	99.9%	57.3%	87.0%
True Model	1	2	3	4	5	6	7	8	9	
True variables number(*)										
3					100	0.8		100	1.5	
11				99.9		99.8	100		100	
21								99.9	1.4	
29							82.0		83.5	
37		100	95.7				100	99.9	99.9	
38			93.6							

Hoover and Perez (1999) used 40 dependent variables as candidates to take the part of the real model, and these variables are correlatively numbered. The number of the variables appearing in the table are the true variables used in the models´ generation.

As in Hoover and Perez (1999) we define: Size=falsely significant variables /(total candidates-possible true variables) and Power = 1-(possible true variables-true variables selected)/possible true variables.

Table 3. Specification searches using Stepwise (at 1% nominal size)

True Model	1	2	3	4	5	6	6A	6B	7	8	9	Means
Category 1	69.4	72.3	68.2	75.1	72.9	1.1	74.7	75.8	71.9	29.5	0.0	55.54
Category 2	30.6	27.3	25.8	24.9	27.1	0.5	25.3	24.2	27.0	11.2	0.0	20.35
Category 3	0.0	0.4	0.0	0.0	0.0	0.0	0.0	0.0	1.1	0.6	0.0	0.19
Category 4	0.0	0.0	3.1	0.0	0.0	27.9	0.0	0.0	0.0	58.6	98.6	17.11
Category 5	0.0	0.0	2.9	0.0	0.0	70.5	0.0	0.0	0.0	0.1	1.4	6.81
Number of true variables	0	1	2	1	1	2	2	2	3	3	5	
Average rate of selected variables	0.37	1.33	2.28	1.30	1.31	1.35	2.30	2.28	3.33	2.61	3.43	
Average rate of true variables	--	1.00	1.94	1.00	1.00	1.02	2.00	2.00	3.00	1.83	3.06	
Average rate of insignificant variables	--	--	--	--	--	--	--	--	--	--	--	
Average rate of falsely significant variables	0.37	0.33	0.34	0.30	0.31	0.34	0.30	0.28	0.33	0.77	0.36	0.37
Type I error (True Size)	0.93%	0.84%	0.89%	0.76%	0.80%	0.88%	0.80%	0.75%	0.89%	2.09%	1.03%	0.97
Power	--	100%	97%	100%	100%	50.8%	100%	100%	100%	60.87%	61.26%	86.99
Frequency true variables included (percent)												
True Model	1	2	3	4	5	6	6A	6B	7	8	9	
True variables number												
3					100	1.6	100	100		100	3.9	
11				100		100	100	100	100		100	
21										41.3	2.4	
29									100		100	
37		100	97.9						100	41.3	100	
38			96.1									

Hoover and Perez (1999) used 40 dependent variables as candidates to take the part of the real model, and these variables are correlatively numbered. The number of the variables appearing in the table are the true variables used in the models´ generation.

As in Hoover and Perez (1999) we define: Size=falsely significant variables /(total candidates-possible true variables) and Power = 1-(possible true variables-true variables selected)/possible true variables.

(1) Except in models 6, and 9, the GASIC improves the identification of true model (Category 1) with respect to HP algorithm. So, as it is possible to observe, the performance of GASIC algorithm is clearly superior with respect to HP. Besides, we didn't eliminate bad specified models, while the HP algorithm does, that is if the general specification fails more than one mis-specification test, the current replication is eliminated and the search begins again with a general specification of a new replication. On average terms, the HP search path has a 46.1% of success meanwhile for GASIC the success increases by up to 63.9 % (excluding models 6A and 6B). With respect to models 6 and 9, both procedures fail considerably in getting the three first categories. (2) For model 6, both procedures (GASIC and HP) provide the same bad results, giving a 0.7% of success in identifying the true model for HP and 0.1% for GASIC. (3) In models 9 the HP algorithm improves the identification of the true model with respect to GASIC because the GASIC never identifies the true model, meanwhile the HP does so 0.8% of the time. Besides, HP is also superior to GASIC in putting the majority of results in category 4. (4) The GASIC algorithm reduces the number of false significant variables, which causes the GASIC to have type I error inferior to HP. So, in model 1 type I error is reduced from 0.7% in HP to 0.17% in the GASIC, in model 2 from 5.7% to 0.19%; in model 3 from 0.9% to 0.26%; in model 4 from 0.8% to 0.24%; in model 5 from 0.7% to 0.19%; in model 6 from 0.6% to 0.18%; in model 7 from 3.0% to 0.22%; in model 8 from 0.9% to 0.18%; and, finally, in model 9 from 3.2% to 0.24%.

Now let's show how our GASIC procedure improves the variable selection provided by the application of Stepwise algorithm at nominal size 1% (the results provided by Stepwise at nominal size 5% are not included because they are worse than nominal size 1%). The results of model identification using Stepwise are shown in Table 3. Comparing these results with those obtained by GASIC in Tables 1 we see that GASIC improves considerably Stepwise, with the only exception of model 6, where the performance of both procedures are equally bad. On average, the GASIC identifies the true model in 68.85% of the time, and Stepwise only does so on 55.54% of the time. Observe that Stepwise tends to over-identify models, where the selection of false significant variables is very high. All of it is corroborated noting the type I error produced in both procedures. So, from a type I error of 0.97% in the Stepwise (the expected would be 1%) a type I error of 0.20% is obtained for the GASIC.

Furthermore, it is possible to compare our results with Hansen (1999). Hansen's selection regressors method starts with the most general model and applies stepwise backward elimination until just a manageable set of K'regressors remains (in his research $K'=10$). At this point, sequentially estimate all 2^{10}models which can be formed from these 10 regressors. For each regression which has been run on this search, calculate and store the SIC. The model with the smallest SIC is the selected model. Hansen (1999) applied this method to HP database and only provided us with the HP category 1, showing the percentage of searches for which the selected specification is the true model (Hansen's Table 1). His results work better than HP's and are similar to our own. Perhaps the small difference could be explained because we don't eliminate any general unrestricted models when mis-specification tests fail like HP (and perhaps Hansen) do.

Nevertheless, it is necessary to point out that Hansen's method is not a general procedure because the number K' of manageable regressors, before sequentially estimating all $2^{K'}$models, is never known in advance. So, it is not safe to suppose, as Hansen did, that $K'=10$. The number of true regressors is crucial in the data mining framework and it is not easy to provide an estimator of k. In the HP database Hansen has no identification problems because the number of independent variables in all models is inferior to 6 and in most models this number is between 1 and 3. So, when Hansen's procedure reduces the number of variable candidates to regressors from 40 down to 10, using Stepwise backward, it is guaranteed with a high probability that the true model contains the dependent variables inside the 10 previously selected variables. Starting with these 10 pre-selected variables, it is possible to estimate all potential $2^{10}=1024$models which are evaluated using SIC criterion. Nevertheless, this procedure becomes useless when the model that we want to identify has a high number of independent variables. For instance if 30 of 60 variables were the number of true variables, and Hansen's method was applied with$K'\geq 30$, it would be necessary to estimate at least $2^{30}=1{,}073{,}741{,}824$ models, which is impracticable. However, GASIC is able to deal with this problem without difficulties. For N=1000 and $R^2=0.50$ the percentage of successful correct retrieval is 44.9%, while for $R^2=0.75$ the percentage of successful correct retrieval is 96.9%. Of course, when both the sample size and the R^2 decrease, these results get worse.

It is not possible to compare directly the percentages of successful retrieval of the 1,2,7, and 9 models by HP, and by GASIC and PcGets in the version of Hendry and Krolzig (1999) given in their Table 2, because they used a sample size T=100 while the HP's sample size data base is 139. It could also be interesting to compare GASIC with the latest versions of PcGets by Hendry and Krolzig (2001, 2003).

We have also introduced a battery of mis-specification tests in the mating process of the GA, in order to discard bad specified models. Nevertheless, at least in the current data base, using mis-specification tests it is only necessary to guarantee the congruence of the general unrestricted model, but it is hardly useful in the process of selecting the correct model proposed by GASIC. So, mis-specification tests don't improve the percentage of successful retrieval of DGPs but, in general, it diminishes the number of generations of GASIC in getting the optimum model.

4.2 Empirical results on several experiments simulated by Perez-Amaral et al. (2003)

Perez-Amaral et al. (2003) proposed a flexible tool for model building based on the Relevant Transformation of the Inputs Network Approach called RETINA. The basic idea of this new procedure is to use out of sample performance measures for model selection. These authors designed several data generating processes and varied several parameters, such as the overall sample size T, the amount of correlation ρ among original variables X and the R^2.

In order to proceed to a direct comparison with our methodology, we are only concerned with the fist four of the seven models simulated by Perez-Amaral et al. (2003). With several transformations, our algorithm is also able to deal with sparse regressors, outliers and structural break. Nevertheless, to keep matters simple, we prefer only to consider the simplest version of our algorithm. So, we compare GASIC capabilities in the following Data Generation Processes (DGPs):

DGP1. *Linear:* $y_i = \alpha_0 + \alpha_1 x_{i1} + \alpha_2 x_{i2} + \sigma u_i \quad i = 1,....,T$, where $\alpha_0 = \alpha_1 = \alpha_2 = 1$, x_{i1} and x_{i2} are jointly normal with correlation ρ between regressors equal to 0.5. The error term u_i is i.i.d. N(0,1) and σ is calibrated to achieve an average R^2 of the resulting estimated equations across replications equal to 0.25, 0.50 and 0.75, respectively.

DGP2. *Ratio:* $y_i = \alpha_0 + \alpha_1 \dfrac{x_{i1}}{x_{i2}} + \sigma u_i \quad i = 1,....,T$, everything else is as in DGP1 except that $\rho = 0.5$ only.

DGP3. *Product:* $y_i = \alpha_0 + \alpha_1 x_{i1} x_{i2} + \sigma u_i \quad i = 1,....,T$ and everything else is as in DGP1.

DGP4. *Linear with binary regressor:* $y_i = \alpha_0 + \alpha_1 x_{i1} + \alpha_3 x_{i3} + \sigma u_i \quad i = 1,....,T$, with $\alpha_3 = 1$ and x_{i3} is a discrete explanatory variable which takes the value 1 with probability 0.5 and 0 otherwise, and everything else is as in DGP1.

As in RETINA for DGP1 to DGP3 we have considered 24 candidate regressors (excluded the constant) constituted by the transformations of the variables included in the DGP (each of which is obtained as $X_{ih}^{\alpha} X_{il}^{\beta}$, $\alpha, \beta = -1, 0, 1$; $h, l = 1, 2, 3$) and of an additional irrelevant variable with the same distribution and correlation as x_{i1} and x_{i2}. For DGP4 we considered 17 candidate regressors for avoiding repetitions of outcomes and divisions by zero. Percentages of successful retrieval of the DGPs by RETINA and by GASIC for different DGPs, sample sizes and $R^2 s$ are reported in Table 4.

As we can see in Table 4, GASIC outperforms RETINA in twenty five out of thirty-five cases, which are signaled by an asterisk in Table 4, and in one case they behaved equally. Several patterns emerge in Table 4. Both procedures are asymptotically similar; that is, when the sample size is high the percentage of successful retrieval of the DGPs are similar. Nevertheless, GASIC outperforms RETINA for small sample sizes and for DGPs with small R^2. Finally, it is necessary to observe that the failures in GASIC retrieving the DGPs are always produced because the GA selects models with the best SIC, although these models are not necessarily the correct models. So, in most cases, the GA works perfectly in the optimization problem and any failure is associated with the SIC as model selection criterion.

Table 4: Percentages of successful retrieval of the DGPs by GASIC and RETINA for different DGPs*, sample sizes and R^2s

DGC	Sample size	$R^2 = 0.25$		$R^2 = 0.5$		$R^2 = 0.75$	
		GASIC	RETINA	GASIC	RETINA	GASIC	RETINA
1: linear	100	10.7	22.8	88.2 *	72.9	94.5	97.7
	200	52.9 *	42.8	97.9 *	93.9	97.0	98.3
	1000	99.6 *	98.6	99.4 *	99.1	99.4 *	99.1
2: ratio	100	79.2 *	39.9	86.2 *	72.6	90.0	93.7
	200	85.3 *	49.8	90.8 *	82.2	94.1	97.4
	1000	94.5 *	73.6	96.5 *	94.7	98.2	99.1
3: product	100	92.2 *	75.9	94.8	96.2	93.6 *	98.6
	200	97.5 *	94.2	97.9	99.1	97.5	99.1
	1000	99.0	99.5	99.4	99.4	99.6 *	99.2
4: linear with binary regressors	100	22.8 *	9.8	80.2 *	43.4	96.3 *	88.1
	200	55.1 *	24.6	98.1 *	72.6	97.9 *	95.5
	1000	99.4 *	89.5	99.7 *	95.7	99.4 *	95.9

The asterisks correspond to the simulations where GASIC outperforms RETINA

5. CONCLUSIONS

In this paper, we have developed the GASIC model selection procedure. The principal contribution of GASIC is a selection of variables method which improves considerably the Stepwise and backward elimination algorithms. It is based on a heuristic optimization method called genetic algorithms which are used to explore the universe of models made available by a general unrestricted model. Once we have checked the congruence of the general unrestricted model using a battery of mis-specification tests, the search process of the true model is only guided by the Schwarz information criterion, which acts as the lost function of the genetic algorithm in order to rank the models. So, GASIC permits exhaustive exploration of the most promising models of the search space without a cumbersome time consuming process.

We strongly agree with Hansen (1999) who emphasized the possibilities of SIC as a power and simple model selection procedure, which is easy to implement in applications, performs at least as well, if not better, than complicated search algorithms. Nevertheless, our paper provides an important improvement in the search algorithm with respect to Hansen's. The drastic reduction of the number of variables practiced in Hansen (1999), who supposes a maximum of ten variables in the model is, in general, an incorrect reduction of the models space and a false assumption. So, GASIC is a speed efficient regressors selector that improves considerably the classic Stepwise or backward elimination.

We have also introduced a battery of mis-specification tests in the mating process of the GA in order to discard bad specified models. Nevertheless, at least in the current data base, using mis-specification tests it is only necessary to guarantee the congruence of the general unrestricted model, but it is hardly useful in the process of selecting the correct model proposed by GASIC. So, mis-specification tests don't improve the percentage of successful retrieval of DGPs but, in general, it diminishes the number of generations of GASIC in getting the optimum model.

We have compared GASIC with recent developments on model selection procedure like Hoover and Perez (1999) based on the general to specific approach of the LSE and like RETINA procedure by Perez-Amaral et al. (2003) who uses out of sample performance measures for model selection. GASIC has a good performance with respect to Hoover and Perez (1999). Comparing GASIC with RETINA, both procedures are asymptotically similar; that is, when the sample size is high, the percentage of successful retrieval of the DGPs is similar. Nevertheless, GASIC usually outperforms RETINA for small sample sizes and for DGPs with small R^2. Furthermore, it could also be interesting to compare GASIC with the latest versions of PcGets like Hendry and Krolzig (2001, 2003).

Besides, our methodology may be employed in selecting correct models with large number of variables on a potentially large set of variables of interest.

Finally, we observe that the most failures of GASIC retrieving the DGPs can only be attributed to the SIC and not to the GA. So, GASIC frequently finds the final models which improve the true model in the sense that they are better behaved with respect to the lost function provided by the SIC. In this sense, the GASIC works well and it is not responsible for the lack of success in getting the correct model. It opens the problem of which could be the suitable lost function in the very badly behaved models of Hoover and Perez (1999) data base.

Lastly, as a final general conclusion, GA opens new perspectives of data mining research which may also be considered in other related fields.

REFERENCES

Akaike, H. (1973): "Information theory and an extension of the maximum likelihood principle", in Petrov, B.N., Csake, F. (Eds), 2nd International Symposium on Information Theory, Budapest, pp. 267-281.

Beenstock, M., Szpiro, G. (2002): "Specification search in nonlinear time-series models using the genetic algorithm". *Journal of Economic Dynamics & Control*, 26, pp. 811-835.

Campos, J., Hendry, D.F., Krolzig, H.-M. (2003): "Consistent model selection by an automatic Gets approach". *Oxford Bulletin of Economics and Statistics*, 65, Supplement, pp. 803-819.

Dorsey, R. D., Mayer, J.M. (1995): "Genetic algoritms for estimation problems with multiple optima, nondiferentiability, and other irregular features". *Journal of Business & Economic Statistics*, 13, pp. 53-66.

Goldberg, D. E. (1989): *Genetic Algorithms in Search, Optimization, and Machine Learning*. Addison Wesley, Reading, Mass.

Hannan, E.J., Quinn, B.G. (1979): "The determination of the order of an autoregression". *Journal of the Royal Statistical Society B*, 41, pp. 190-195.

Hansen B. E. (1999). "Discussion of Data mining reconsidered". *Econometrics Journal*, 2, pp. 192-201.

Hendry, D. F., Krolzig, H.-M. (1999): "Improving on "Data mining reconsidered" by K.D. Hoover and S.J. Perez". *Econometrics Journal*, 2, pp. 202-219.

Hendry, D. F., Krolzig, H.-M. (2001): *Automatic Econometric Model Selection Using PcGets 1.0*, Timberlake Consultants Press, London.

Hendry, D.F. and Krolzig, H.-M. (2003): *The Properties of Automatic Gets Modelling. Economics Department*. Oxford University.

Holland, J. (1975): *Adaptation in Natural and Artificial Systems*. Ann Arbor. The University of Michigan Press.

Hoover, K. D., Perez, S.J. (1999): "Data mining reconsidered: encompassing and the general- to-specific approach to specification search". *Econometrics Journal*, 2, pp. 167-191.

Koza, J. (1992): Genetic Programming. MIT Press, Cambridge, MA.

Krolzig, H.-M., Hendry, D. F. (2001): "Computer automation of general-to-specific model selection procedures". *Journal of Economic Dynamic & Control*, 25, pp. 831-866.

Lovell, M.C. (1983): "Data mining". *Review of Economics and Statistics*, 65, pp. 1-12.

Perez-Amaral, T. Gallo, G. M., White, H. (2003): "A Flexible Tool for Model Building: the Relevant Transformation of the Inputs Network Approach (RETINA)". *Oxford Bulletin of Economics and Statistics*, 65, Supplement, pp. 821-838.

Mallows, C.L. (1973): "Some Comments on C_p". *Technometrics*, November, pp. 661-676.

Miller, A. (2002): *Subset Selection in Regression*. Chapman & Hall/CRC.London.

Schwarz, G. (1978): Estimating the dimension of a model. *Annals of Statistics*, 6, pp. 461-464.