Original paper

Soft Computing 8 (2004) 527–533 © Springer-Verlag 2003
DOI 10.1007/s00500-003-0310-2

Genetic algorithms for outlier detection and variable selection in linear regression models

J. Tolvi

Abstract This article addresses some problems in outlier detection and variable selection in linear regression models. First, in outlier detection there are problems known as smearing and masking. Smearing means that one outlier makes another, non-outlier observation appear as an outlier, and masking that one outlier prevents another one from being detected. Detecting outliers one by one may therefore give misleading results. In this article a genetic algorithm is presented which considers different possible groupings of the data into outlier and non-outlier observations. In this way all outliers are detected at the same time. Second, it is known that outlier detection and variable selection can influence each other, and that different results may be obtained, depending on the order in which these two tasks are performed. It may therefore be useful to consider these tasks simultaneously, and a genetic algorithm for a simultaneous outlier detection and variable selection is suggested. Two real data sets are used to illustrate the algorithms, which are shown to work well. In addition, the scalability of the algorithms is considered with an experiment using generated data.

Keywords Variable selection, Model selection, Outlier, Outlier detection, Genetic algorithm

1 Introduction

In statistical data analysis, outliers or aberrant observations are, according to one definition, observations that are somehow different from the majority of the data. This definition is somewhat vague, but it will be adequate for our purposes in this article. There are several statistical methods for outlier detection in different circumstances. In practical work it may however be difficult to decide which method to use, and to resolve potential inconsistencies in the results obtained by different methods. And if outliers are detected in the data, there are several ways of taking them into account in the analysis. For example, one can either remove the outlying observations from the data altogether, or incorporate the detected outliers into the statistical model. An introduction to outliers, and their detection and modeling can be found in [2].

In addition to the basic problems of outlier detection mentioned above, there is an additional nuisance in practical outlier detection, namely the possibility of smearing and masking. These terms refer to two related special problems. First, smearing means that an outlier causes another observation, which is not in reality an outlier at all, to be considered as one by an outlier detection method. Second, masking is said to occur when an outlier prevents another one from being detected by an outlier detection method. Sequential detection of outliers may therefore be misleading, if the detection of one outlier causes the subsequent detection of other outliers to be flawed, due to either smearing or masking, or even both.

In this article therefore a simultaneous outlier detection method is first considered in linear regression modeling. The simplest outlier detection method would be to consider all possible outlier combinations in turn, that is to go through all possible permutations of the observations into two groups, outliers and non-outliers, and decide which of these is the best combination (based on some prespecified criterion). This, however, becomes impossible in practice due to the enormous amount of possible combinations ($2^N - 1$, where N is the number of observations in the data). Since some combinations are more probable than others, a genetic algorithm (GA) is a natural way of going through the most interesting possibilities. The same idea has recently also been used in [1] in detecting outliers from time series data.

The use of GAs for variable selection in statistical modeling has already been discussed in, for example, [4] and [16]. In addition to outlier detection, we will extend our algorithm to include also a choice on which variables to select, out of a set of candidate variables, into the regression model. The motivation for this is that the choice of which variables to select into the model can affect the outlier detection, and conversely, the choice of whether or not to consider some observations as outliers can affect the variable selection [3]. If the choices on outliers and variables are made sequentially, it is possible that mistaken conclusions are made in either, or even in both steps. A simultaneous choice, as in our algorithm, should be more appropriate in this respect.

The fitness functions for these GAs are built on using information criteria. They are used both to select the variables of the model, and to determine the outlying observations. The idea of using information criteria to

Published online: 7 October 2003

J. Tolvi
Department of Economics, University of Turku,
20014 Turku, Finland
e-mail: jussi.tolvi@utu.fi

I would like to thank Dr Tero Aittokallio and an anonymous referee for useful comments.

detect outliers was apparently first introduced in [12], in finding outliers in a sample of independently and identically distributed data. A general discussion of model selection using information criteria is given in [8], and [17] discusses their use in variable selection in autoregressive moving average (ARMA) time series models.

GAs have been used for somewhat similar purposes before, and we will give just a few examples here. First, [5] presents a GA for model selection of subset ARMA models. If such models are considered for a time series, there exists a large number of potential model forms, and finding the best one using conventional methods may require the estimation of a prohibitive number of models. [11] describes a GA for the detection of multiple change-points (or level shifts) in a time series. The problems caused by change-points are similar to those caused by outliers, in that the detection of one change-point affects the consequent detection of further change-points. A simultaneous, rather than a sequential, detection procedure is therefore beneficial also for change-points. Finally, a related area of research is that of the use of GAs for feature and instance selection in, for example, data mining. A number of references to this literature can be found in [10].

2 Models and outliers

To begin with, we will consider only outlier detection in linear regression models, when the variables to be included in the model are known. The observed data consists of the dependent variable vector $\mathbf{y}$, and of the independent explanatory variable vectors $\mathbf{x}_1, \ldots, \mathbf{x}_p$. We will use the notation z_i for an element of a vector $\mathbf{z}$. In addition, $\mathbf{X}$ will denote a matrix of the column vectors $\mathbf{x}_1, \ldots, \mathbf{x}_p$. The linear regression model can now be written as

$$\mathbf{y} = \mathbf{X}\beta + \boldsymbol{\varepsilon} \ , \qquad (1)$$

where $\mathbf{y}$ is an $N \times 1$ vector of the dependent variable, $\mathbf{X}$ the $N \times (p+1)$ matrix of independent variable vectors, including a constant vector $\mathbf{1}_{(N\times 1)}$, β a $(p+1) \times 1$ parameter vector and ε an $N \times 1$ vector of independent and identically distributed Gaussian errors, $\varepsilon_i \sim N(0, \sigma^2)$, $i = 1, \ldots, N$. The model can be estimated by ordinary least squares regression, which provides estimates for the parameters ($\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$) and residuals ($\hat{\varepsilon} = \mathbf{y} - \mathbf{X}\,\hat{\beta}$).

If outliers occur in the data, the errors can be thought to have a distribution different from the above. There are several possibilities, but perhaps the most intuitive one is the mixture model, where with probability $1 - \pi$ each error ε_i comes from a $N(0, \sigma^2)$ distribution, and with probability π from another distribution. Here, $0 < \pi < 1$, and if outliers are thought of as rare events, as is usually meaningful, π must be close to zero. There are also various alternatives for the outlier distribution. For example, the outliers may come from a $N(\omega, \sigma^2)$ distribution, or from a $N(0, \delta\sigma^2)$ distribution, where ω and δ are some fixed constants ($\omega \neq 0, \delta > 1$). In either case, some errors ε_i, and consequently some observations y_i are somehow different from the majority of the data.

The detection of outliers is important, not only for their own sake, but also because the inferences drawn from the model will be biased if outliers are neglected. Potential outliers can be incorporated into the linear regression model of equation (1) by the use of dummy (or indicator) variables. A dummy variable is an $N \times 1$ vector that has a value of one for the outlier observation, and zero for all other observations. For example, a dummy variable to be added to the model above could be $\mathbf{x}_{p+2} = (1 \quad 0 \quad \cdots \quad 0)'$. This dummy variable would correspond with the first observation being an outlier. A dummy variable in the regression model is therefore equivalent to a detected outlier, and the problem here is to select the best model, where the candidate models have different combinations of all possible dummy variables (i.e. corresponding to each N observations) as explanatory variables.

Outlier detection, and later also variable selection, are in this article based on the use of information criteria. There is a large number of possible criteria to choose from. The BIC criterion of [18] will be used here, since it usually performs quite well in various situations, see for example [15]. For our linear regression model with dummy variables the criterion can be calculated as

$$\mathrm{BIC} = \log(\hat{\sigma}^2) + m\log(N)/N \ , \qquad (2)$$

where N is the sample size, $\hat{\sigma}^2 = (\hat{\varepsilon}'\hat{\varepsilon})/N$ is the residual sum of squares, and $m = 1 + p + m_d$, the total number of parameters in the estimated model, consists of parameters for the constant, the p independent variables and the number of outlier dummies m_d. In general, a good model has small residuals, and few parameters. A model with the smallest value of BIC is therefore preferred.

A problem in using the BIC for outlier detection is that by itself it tends to include unnecessary outlier dummies (see also [6]). To circumvent this problem, a correction to the criterion will be used in this article. The corrected BIC takes into account the different nature of outlier dummies and other variables, and has a different penalty term for different variables. This takes the form of an extra penalty for the dummies. The corrected BIC, denoted BIC′, is given by

$$\mathrm{BIC}' = \log(\hat{\sigma}^2) + (1+p)\log(N)/N + \kappa m_d \log(N)/N \ , \qquad (3)$$

where κ ($\kappa > 1$) is the extra penalty given to outlier dummies. Simulation experiments can be conducted to determine relevant values for κ in different situations. One can, for example, generate artificial data from the statistical model, estimate the model both with and without outlier dummies, and find a value of κ with which an outlier dummy is unnecessarily accepted into the model with some small probability, say one per cent. We have found that often a value of three works well in preventing unneeded dummies from being added to the model, yet at the same time ensuring that true outliers are detected, and a dummy variable added for them. This value ($\kappa = 3$) will be used in the examples reported later in the article.

3 A genetic algorithm for outlier detection

A general introduction to genetic algorithms is [7]. In the application of this article the two most important

considerations are the coding of the models and the fitness function. In addition, several algorithm parameters, and procedures for the creation of new generations from the old ones, must be decided on. In a GA the best solution to a certain problem, such as optimization of a function or a combinatorial problem, is found by considering a large number of possible solutions, and combining aspects of potentially good ones. In our application the problem is to describe the data by finding the best possible model, first by finding any outliers in the data. To this end a number of, initially randomly selected, candidate models are first examined. The goodness of each model is determined by the fitness function. The next generation of candidate models are formed from the current models, using a process of combination described below. When a large number of generations has been examined, the GA should discover, if not the overall best possible model, at least a reasonably good one.

We will start by describing an algorithm for outlier detection. The coding of the candidate models for outlier detection is straightforward. Each model, also called an individual, is fully described by a binary vector $\mathbf{z} = (z_1 \quad \cdots \quad z_N)$, where $z_i = 0$ indicates no outlier dummy and $z_i = 1$ indicates an outlier dummy for observation i, for each $i = 1, \ldots, N$. These elements of the $\mathbf{z}$ vector are also referred to as the genes of that individual. For example, a model with a dummy variable for the first and last observations is described by the vector $\mathbf{z} = (1 \quad 0 \quad \cdots \quad 0 \quad 1)$. Note that before the GA is run on a data set, the dummy variables for each observation must first be created. The fitness of an individual is then computed simply as the BIC′ value (3) for a linear regression model (1) with the corresponding dummy variables.

In our GA, the population size in each generation is 40 individuals. The initial population for the algorithm to start with is generated randomly, such that each gene of each individual has a value of zero with probability 0.9, and therefore a value of one with probability 0.1. Linear regression models corresponding to these individuals are then estimated using the observed data, and BIC′ values for them computed. Note that the algorithm could most likely be made considerably faster, by incorporating any preliminary information about which of the observations are potentially outliers. Such information could come from, for example, knowledge of the data collection, or simply by looking at the data. Here no such information has been used.

Mating, or the creation of the next generation of individuals from the previous one, is also based on the BIC′ values of the individuals. Each new individual of generation $t + 1$ has as its parents two individuals from the previous generation t. These are selected by first computing for each individual j in generation t a uniform random variable in the interval $[0, 1/\mathrm{BIC}'_j]$.[1] The two individuals with the largest values of this random variable are then selected as parents. In this way the best individuals, that is the ones with the smallest values of the fitness function BIC′, are more likely to pass their genes onto the next generation. To create the offspring, two integer points, ν_1 and ν_2, are first randomly selected from the interval $[0, N - 1]$, such that $\nu_2 \geq \nu_1$. The offspring gets the first ν_1 genes from the first parent, the next $\nu_2 - \nu_1$ genes from the second parent, and the last $N - \nu_2$ genes again from the first parent. Note that if $\nu_1 = \nu_2$, the offspring is an exact copy of the first parent. This procedure is repeated to create the same number of individuals as existed in the previous generation.

To make sure that the algorithm will also take into consideration entirely new models, and that as many models as possible are examined, mating of the individuals from the previous generation will not be enough. In evolutionary terms, more genetic variation in the population is needed. To this end, the individuals of each generation are also mutated before model estimation. Each gene of each individual is flipped, from zero to one or vice versa, with probability 0.01. After the mating and mutation steps are completed, models are estimated and BIC′ values computed for the new generation $(t + 1)$, and a new round of mating, mutation and estimation is begun on the next generation $(t + 2)$. Table 1 has the most important GA parameters.

In addition to mating and mutation, a condition for the maximum number of dummies is used to alter the population. This rule is used in order to keep the candidate models from having too many variables, since only a few dummies will presumably be allowed in the final model. The rule states that if a candidate model has more than $N/5$ dummy variables, or in other words if the number of outliers is more than 20% of the number of observations, it is dropped from consideration.

The mating, mutation and estimation procedure outlined above is repeated until convergence is achieved. The convergence criterion states that when no improvement is found in the last M generations, the algorithm is terminated, and the best model so far found is reported. The value of M has to be chosen for each application of the algorithm. A large enough value must be chosen to make sure that a good solution is found, but on the other hand a larger value leads to more computations and a longer execution time.

After some experiments with the algorithm, it was noted that the results can be improved if a small number of the best individuals are kept the same from one generation to the next. It seems that by preserving the two best individuals in each generation intact, the time needed to find the optimal solution can be cut by as much as half. In the remainder of this article, this strategy is therefore always followed.

Since the algorithm is based on a random choice of models to be considered, different results can be obtained in a run with different random numbers. Therefore, and also to find appropriate values of the termination criterion for each application, the algorithm was run several times, using different seeds for the random number generator each time.

[1] Where BIC'_j denotes the BIC′ value of individual j. This interval can be used in this situation since the BIC′ values are all larger than one. If this were not the case, some other transformation of the BIC′ values would have to be used in selecting the parents. This is what happens later in Sect. 7, where generated data is used, and a constant has to be added to the BIC′ values before mating.

Table 1. Information on the data sets used, GA parameters and results

	Scottish hill running data	Stack loss data
Observations	35	21
Outliers	3	4
Explanatory variables[a]	2	3
Variables that belong to the model[a]	2	2
Crossover probability	1	1
Mutation probability	0.01	0.01
Population size	40	40
Elite population size[b]	2	2
Outlier detection only		
Number of runs	10	10
Minimum[c]	7	21
Maximum[c]	38	214
Average[c]	18	97
Number of times best model found	10	10
Simultaneous variable selection and outlier detection		
Number of runs	10	10
Minimum[c]	10	170
Maximum[c]	68	1890
Average[c]	34	829
Number of times best model found	10	10

[a] Excluding the constant
[b] Crossover and mutation operators were not used on these individuals.
[c] Number of generations needed to find the best model

4 Examples

Our empirical examples use two data sets that have been used in several earlier papers. References to these, and other further information, including where to obtain the data can be found in [9].[2] These data sets have specifically been used to illustrate outlier detection and variable selection methods in linear regression modeling. It is therefore interesting to see how our genetic algorithms do in comparison with earlier findings. In this section we will detect outliers from these data sets with our GA. In a later section we will use a GA also to select the variables to include in the statistical models. Some information on the data sets can be found in Table 1.

The first data set, called the Scottish hill running data, has 35 observations. The dependent variable is the record times (in minutes) of hill running races, and the two independent variables are distance (length of the race in miles) and climb (the total elevation gained in the race in feet). In earlier research observations 7 and 18 have been initially identified as outliers.[3] If these two observations are removed from the data, it is usually found that observation 33 seems also to be an outlier. Observations 7 and 18 therefore mask observation 33 from being detected as an outlier. At this stage, our statistical model for this data includes a constant and both of the independent variables. Our only task is therefore to detect any possible outliers in the data.

The GA described earlier was run with this data, using ten different seeds for the random number generator. All runs result in the same outliers being detected, at observations 7, 18 and 33. These three observations are exactly the ones identified as outliers also in the earlier work mentioned above. The minimum, maximum and average number of generations needed to find this best solution are given in Table 1. The solution was always found quickly by the GA, since the average number of generations needed was only 18. The estimated model with the three outlier dummies has a BIC′ value of 4.13599.

The second data set, called the stack loss data, has 21 observations from a chemical process. The dependent variable (stack loss) is the amount of the chemical escaping the plant. The three independent variables are air flow, cooling water temperature and concentration of acid used in the process. Observations 1, 3, 4 and 21 have in earlier research been identified as outliers. This data set provides an interesting and rather extreme example of masking. It is noted in [9] that the detection of any of these outliers is very difficult if only one observation at a time is examined, but that simultaneous methods are able to detect all four outliers.

The GA was run with this data as well, with a model that has a constant and all three variables. Again, ten runs were made with different seeds for the random number generator, and the results can be found in Table 1. The best outlier combination found had observations 1, 3, 4 and 21 as outliers, with an estimated BIC′ value of 2.29069. This solution, which was found in all ten runs, is therefore again the same that has been found in earlier work as well. This solution was found, on average, in less than 100 generations.

However, and as noted above, the detection of the outliers in this data set is indeed rather difficult. If only two outliers are in the current model (observations 4 and 21 are fairly easy to detect as outliers), adding either of the two remaining ones alone will increase the BIC′ value. Therefore moving from a model with two dummies to the optimal model with four dummies can be rather difficult. This illustrates the need to either run the GA several times with different random number seeds, or set the termination criterion such that a larger number of generations are considered before termination. Here, it would seem that the algorithm can be safely terminated if no improvement is found in a few hundred generations.

Since the sample sizes in our examples are so small, it was feasible to consider, as a comparison, all possible outlier combinations. This amounts to, in other words, finding the best solution by enumerating all possibilities. The Scottish hill running data has more observations, and was chosen for this purpose as the more demanding example. Models for all possible outlier combinations with up to seven outliers were considered. The best discovered outlier combination was the same that was found with the

[2] At the time of writing, the data were available from one of the authors' website. See `http://www.stat.colostate.edu/~jah/index.html`.
[3] It is pointed out in [9] that observation 18 has a recording error in the independent variable. All other observations are apparently correct, but a few of them are aberrant (i.e. outliers).

genetic algorithm, that is that observations 7, 18 and 33 are outliers. Estimating all these models, and calculating their BIC′ values took approximately four hours. This can be compared with the few seconds needed to run the GA for several hundred generations, by which time the best solution has certainly been discovered.

The role of the penalty function was also examined further. The robustness of the outlier detection results with respect to the fitness function was tested by running the algorithm for the Scottish hill running data with different values for κ, the extra penalty given to outlier dummies. In addition to the previously used value of 3, the algorithm was also run with values of 2, 2.5, 3.5, and 4. The same outliers as before were detected with all these values, apart from the value 2, with which an additional outlier at observation 19 was detected. The obtained results would therefore have been the same with a wide range of penalty function values.

5
Simultaneous variable selection and outlier detection

In the previous sections a genetic algorithm was used to detect outliers in a situation where the variables included in the statistical model had already been chosen. For such situations there exists a large number of different outlier detection methods, and the use of the proposed genetic algorithm method may not perhaps be the best possibility.

For a more useful application, outlier detection can be combined into a more general model selection algorithm. The logical next step in this direction is a GA for simultaneous variable selection and outlier detection. Similar work has already been carried out using other methods. For example, [9] considers the simultaneous variable selection and outlier detection in linear regression models in a Bayesian framework. This method, however, requires the analyst to first identify a subset of observations that are potentially outliers. The method then selects the final outliers from this group. In our GA no such preliminary work is needed. Similarly, [13] and [14] employ Bayesian methods to examine whether some macroeconomic time series are better characterized by nonlinear models, or by linear models that contain structural changes and outliers. In both these situations the questions of outlier detection and model selection are related.

It is clear that the choices on the statistical model and the variables to include in the model influence the results of outlier detection, since any observation is an outlier only in relation to some specific model. On the other hand the choice of whether to correct some outliers might also influence the outcome of model and variable selection. The traditional procedure is to first select the statistical model and variables, and then detect possible outliers within the chosen model. This sequential modeling may lead to conflicting outcomes, depending on the order in which the tasks are carried out. If, on the other hand, the choice on the variables to include in the model, and the choice on which observations to consider as outliers is made simultaneously, these problems can perhaps be avoided. We propose in this section one possible solution to this using a GA similar to the one described earlier for outlier detection.

The extension of the outlier detection GA to include also variable selection is straightforward. Each model, consisting of the variables included in the statistical model and outliers detected, is described by a binary vector $\mathbf{z} = (z_1 \quad \ldots \quad z_{p+1+N})$, where the first $p+1$ elements indicate whether the model contains the corresponding column of the $\mathbf{X}$ matrix as an explanatory variable, and the final N elements indicate whether or not an outlier occurs at observation $i = 1, \ldots, N$, and consequently whether a dummy variable is added for that observation. Otherwise the same algorithm specifications can be used as above.

6
Examples continued

The GA presented in the previous section was next run on the two data sets, this time including the choice of the variables to include in the model in the algorithm as well. The algorithm was again run ten times for both data sets, with different seeds for the random number generator. The results are again given in Table 1.

For the Scottish hill running data, the best estimated model had a BIC′ value of 4.13599. All of the three variables (the constant and both explanatory variables) were included in the best model, as in [9]. And as earlier, observations 7, 18 and 33 were again detected as outliers. The selected model is therefore exactly the same as the one found earlier in section 4. Here the best model was, on average, discovered after about 30 generations. This number is not much more than what was required for the outlier detection alone.

For the stack loss data set, the third variable, acid concentration, does not belong in the model according to [9]. For this data, the best estimated model had a BIC′ value of 2.23295, and includes a constant and the first two explanatory variables, and detects the same outliers (at observations 1, 3, 4, and 21) as earlier. The third explanatory variable is left out of the model. These findings are again in line with earlier research. Note also that the BIC′ value obtained earlier in the outlier detection algorithm is indeed larger than the one obtained here, indicating that dropping the third variable improves the model fit based on the BIC′ criterion. The average number of generations needed to find this model is over 800. This is mainly due to the difficulties with masking noted earlier: all four outliers must be discovered at the same time. This also means that the stopping criterion has to allow more time for the best solution to be discovered, and several thousand generations without an improvement should probably be iterated before stopping.

7
Scalability of the algorithm

Both of the two data sets used earlier as examples are rather small, both in terms of the number of observations and the number of variables in the data. To get some idea on the scalability of these algorithms, some experiments were therefore conducted by examining generated data sets of different sizes. In this section we will concentrate on the simultaneous variable selection and outlier detection algorithm, since this task is clearly more demanding than outlier detection alone. The generated data has two

Table 2. Scalability of the simultaneous outlier detection and model selection algorithm – experimental results on generated data[a]

$N(n)$	50 (4)	50 (8)	50 (12)	100 (4)	100 (8)	100 (12)	200 (4)	200 (8)	200 (12)	300 (4)	300 (8)	300 (12)
Minimum[b]	3	10	21	13	46	32	128	103	230	1002	1593	1596
Maximum[b]	16	46	68	57	141	135	1221	1044	942	2686	2994	5210
Average[b]	8	23	46	31	78	81	346	475	500	1785	2289	2480
Time[c]	1	3	6	13	32	35	558	791	850	6193	8149	9038
Generations/s[d]	7.7	7.4	7.2	2.5	2.4	2.3	0.6	0.6	0.6	0.3	0.3	0.3

[a] GA parameters are as in Table 1. N is the number of observations, n the total number of explanatory variables in the data set.
[b] Number of generations needed to find the true model.
[c] Total computation time needed for 10 runs, relative to the simplest case of $N = 50, n = 4$.
[d] Average number of generations iterated per second.

kinds of variables, those that belong in the true model and those that do not. It also has a certain number of outliers, as detailed below. The GA is run on the generated data, until the true model is found by the GA as the best model so far. The algorithm then terminates and the number of generations iterated so far is reported. Ten runs with different random number generator seeds are again executed for each combination of the factors.

The total number of explanatory ($\mathbf{x}$) variables, n, in the data set is varied, such that $n = 4, 8, 12$. The sample sizes used are $N = 50, 100, 200, 300$. The data is generated as follows. First, the explanatory variables are generated from the standard Gaussian distribution. Half (the first $n/2$) of these are then used to generate the independent ($\mathbf{y}$) variable, along with random Gaussian error terms ε, such that $\mathbf{y} = \mathbf{x}_1 + \mathbf{x}_2 + \ldots + \mathbf{x}_{n/2} + \varepsilon$. In other words, half of the explanatory variables in the data set always belong to the true model. Note that to simplify matters slightly, no constant variables are used in this experiment, either in the data or in the estimated models.

Outliers are then added to this data, such that 2% of all observations are outliers, or that the probability of an outlier occurring is 0.02 for every observation. In addition, the outliers' size, ω, is also increased as the sample size increases. This is not an unrealistic assumption, since in larger samples more extreme observations appear due to random variation. The outlier size ω is 4 for $N = 50$, 5 for $N = 100$, 7 for $N = 200$ and 8 for $N = 300$. To add outliers to the data, a value of ω is simply added to the first 2% of the $\mathbf{y}$ variable observations, so that $y_i = y_i + \omega, i = 1, 2, \ldots, N/50$ (which is identical to drawing the corresponding errors from a Gaussian distribution with a mean of ω). The actual number of outliers in the data sets varies therefore from one ($N = 50$) to six ($N = 300$).

Table 2 has the results of this experiment. The minimum, maximum and average numbers of generations (from the 10 runs) needed to find the true model are first reported. As can be expected, the computational burden in terms of the number of generations needed to find the true model is increased both by an increase in the number of variables, and by an increase in the sample size. The sample size seems to be the more important factor in this respect, however, since increasing the sample size from 100 to 200 to 300 will increase the average number of generations needed to find the true model from tens to hundreds to thousands.

The number of generations needed to find the true models does not tell the whole story of course, since models with more variables and data sets with more observations require also more computational time for each model and therefore each generation. The total computation time needed for all 10 runs (relative to the smallest data set of $N = 50, n = 4$) and the average number of generations iterated per second are therefore also reported in Table 2, to give some idea of the true computational costs.[4] The total computational time is more or less insignificant for samples of 50 and 100 observations. For a sample of 200 observations, however, the times are measured in minutes rather than seconds, and for a sample of 300 observations in hours rather than minutes. The number of generations iterated per second decreases clearly as the sample size grows, but decreases only very slightly as the number of variables is increased.

One must also keep in mind that with real data the true model is of course not known, and therefore actual computational times would be longer, since a large number of generations have to be iterated without improvement before stopping the algorithm. Overall, it seems therefore that with present-day computers the proposed method is practicable, but only for relatively small sample sizes.

8 Discussion

Genetic algorithms for outlier detection and variable selection in linear regression models were presented in this article. It seems that the GAs are able to avoid the potential problems of smearing and masking, which sometimes cause problems for reliable outlier detection. In addition, simultaneous outlier detection and variable selection is possible.

Although this article considered only linear regression models, the idea of using GAs for outlier detection and variable selection can be applied in several statistical models. The discussion here was in the context of cross-section data, but the algorithms could obviously be used also for time series data. In addition to variable selection and outlier detection, similar GAs could be built to simultaneously consider also other kinds of modeling choices. These could include, for example, the detection of change-points in time series models, as in [11].

In future work we intend to use similar GAs for modeling time series of industrial production, which are known to have some kind of nonlinearity [13, 14]. The precise form of

[4] A relatively slow Pentium PC (233MHz) was used for all computations.

the nonlinearity is not obvious, however, and therefore selection between alternative models, such as different linear and nonlinear models, both with and without dummy variables for level shifts and outliers, using a GA could provide interesting results. Unfortunately, however, in many cases the estimation of the models is considerably more costly in CPU time, which perhaps limits the practical applicability of the method somewhat.

One further point, not deemed to be of sufficient interest to be tackled here, is the fact that exactly the same models are estimated several times during the run of the GA. One possibility to circumvent this is to keep track of all the models that have already been considered, and estimate each model only once, as in [11]. If the estimation of the models were more time-consuming, this would clearly be worth considering. Other similar practical improvements to the algorithm and the required computations are also surely possible.

References

1. **Baragona R, Battaglia F, Calzini C** (2001) Genetic algorithms for the identification of additive and innovation outliers in time series. Comput Stat Data Anal **37**: 1–12
2. **Barnett V, Lewis T** (1994) Outliers in Statistical Data. 3rd edn John Wiley, Chichester
3. **Chatterjee S, Hadi AS** (1988) Sensitivity Analysis in Linear Regression. Wiley, New York
4. **Chatterjee S, Laudato M, Lynch LA** (1996) Genetic algorithms and their statistical applications: an introduction. Comput Stat Data Anal **22**: 633–651
5. **Gaetan C** (2000) Subset ARMA model identification using genetic algorithms. J Time Series Anal **21**: 559–570
6. **George EI** The variable selection problem (2000) J Amer Stat Assoc **95**: 1304–1308
7. **Goldberg DE** (1989) Genetic Algorithms in Search, Optimization, and Machine Learning. Addison-Wesley, Reading, MA
8. **Granger CWJ, King ML, White H** (1995) Comments on testing economic theories and the use of model selection criteria. J Econometrics **67**: 173–187
9. **Hoeting J, Raftery AE, Madigan D** (1996) A method for simultaneous variable selection and outlier identification in linear regression. Comput Stat Data Anal **22**: 251–270
10. **Ishibuchi H, Nakashima T, Nii M** (2001) Genetic-algorithm-based instance and feature selection. In: Liu H, Motoda H (eds), Instance Selection and Construction for Data Mining. Kluwer
11. **Jann A** (2000) Multiple change-point detection with a genetic algorithm. Soft Comput **4**: 68–75
12. **Kitagawa G** (1979) On the use of AIC for the detection of outliers. Technometrics **21**: 193–199 Corrigenda: Technometrics, **23**: 320–321
13. **Koop G, Potter S** (2000) Nonlinearity, structural breaks or outliers in economic time series? In: Barnett WA, Hendry DF, Hylleberg S, Teräsvirta T, Tjøstheim D, Würtz A (eds), Nonlinear Econometric Modeling in Time Series. Cambridge University Press, Cambridge
14. **Koop G, Potter S** (2001) Are apparent findings of nonlinearity due to structural instability in economic time series? Econometrics J **4**: 37–55
15. **Mills JA, Prasad K** (1992) A comparison of model selection criteria. Econometric Rev **11**: 201–233
16. **Minerva T, Paterlini S** (2002) Evolutionary approaches for statistical modelling. In: Proceedings of the 2002 Congress on Evolutionary Computation, Vol 2, pp. 2023–2028
17. **Pötscher BM, Srinivasan S** (1994) A comparison of order estimation procedures for ARMA models. Statistica Sinica **4**: 29–50
18. **Schwarz G** (1978) Estimating the dimension of a model. The Annals Stat **6**: 461–464