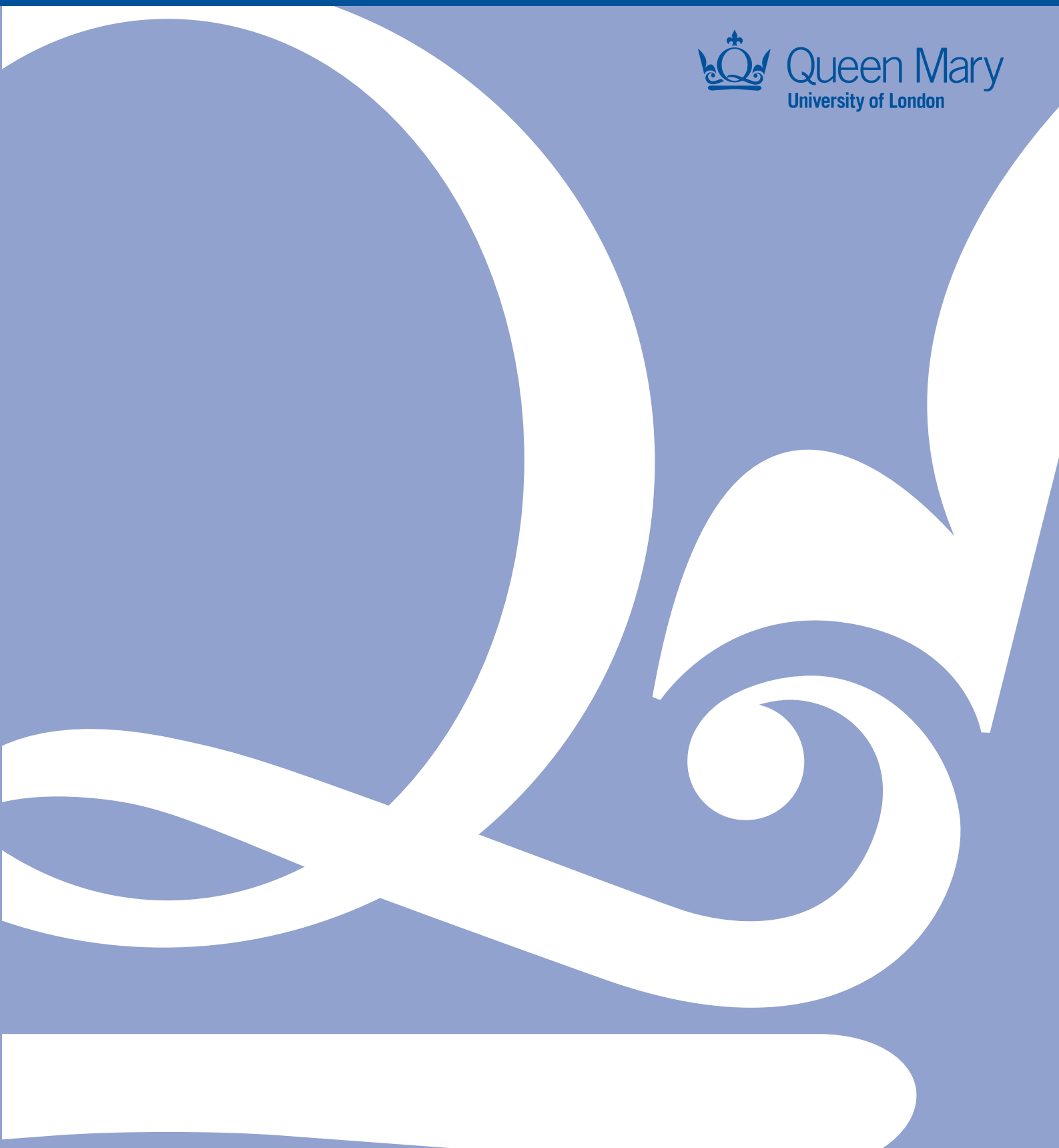
Department of Economics

Choosing the Optimal Set of Instruments from Large Instrument Sets

George Kapetanios

Working Paper No. 534 May 2005 ISSN 1473-0278



Choosing the Optimal Set of Instruments from Large Instrument Sets

George Kapetanios*
Queen Mary, University of London

March 4, 2005

Abstract

It is well known that instrumental variables (IV) estimation is sensitive to the choice of instruments both in small samples and asymptotically. Recently, Donald and Newey (2001) suggested a simple method for choosing the instrument set. The method involves minimising the approximate mean square error (MSE) of a given IV estimator where the MSE is obtained using refined asymptotic theory. An issue with the work of Donald and Newey (2001) is the fact that when considering large sets of valid instruments, it is not clear how to order the instruments in order to choose which ones ought to be included in the estimation. The present paper provides a possible solution to the problem using nonstandard optimisation algorithms. The properties of the algorithms are discussed. A Monte Carlo study illustrates the potential of the new method.

Keywords: Instrumental Variables, MSE, Simulated Annealing, Genetic Algorithms

JEL Codes: C12, C15, C23

1 Introduction

It is well known that instrumental variables (IV) estimation is sensitive to the choice of instruments both in small samples and asymptotically. Asymptotic efficiency is obtained by using all valid available instruments but finite sample performance of IV estimation need not be optimal for this choice. (see, e.g., Morimune (1983) or Bound, Jaeger, and Baker (1996)).

Recently, Donald and Newey (2001) suggested a simple method for choosing the instrument set. The method involves minimising the approximate mean square error (MSE) of a given IV estimator where the MSE is obtained using refined asymptotic theory. In particular, Donald and Newey (2001) use expansions similar to those suggested by, e.g., Nagar (1959)

*Department of Economics, Queen Mary, University of London, Mile End Road, London E1 4NS. email: G.Kapetanios@qmul.ac.uk

and Rothenberg (1983), to provide expressions for the approximate MSE of standard IV estimators such as two-stage least squares (2SLS) and limited information maximum likelihood (LIML). The expansions are provided for the case where the number of instruments grows with the sample size, n, but at a slower rate than n. The problem addressed by Donald and Newey (2001) is one of two separate but related problems concerned with instrument selection. It relates to choosing a subset of valid instruments that minimises MSE for a particular IV estimator and is therefore designed to improve the performance of a consistent estimator. A related problem is that addressed by Andrews (1999) where criteria, similar in spirit to information criteria, are used to select the largest possible set of valid instruments (or more generally moment conditions) among a set of possibly valid instruments. The methods we discuss in this paper may be easily adapted to this distinct problem.

An issue with the work of Donald and Newey (2001) is the fact that when considering large sets of valid instruments, it is not clear how to order the instruments in order to choose which ones ought to be included in the estimation. When a researcher has N potential instruments, then there exist 2^N possible sets of instruments to be considered. Strictly speaking, one needs to compute the MSE of all these sets before choosing the optimal one than minimises MSE. Clearly, even for moderate N such as, say, $N = 20$ or $N = 30$, this is a formidable computational task. Furthermore, as Donald and Newey (2001, pp. 1164) point out such a search is not recommended as it is likely to lead to an estimator of the optimal set which is too variable.

In some cases an ordering of the instruments may be possible following economic theory. But in a large number of cases no such ordering may be possible. Even if some instrument is more useful in estimation than some other instrument, one needs to know the identity of the instrument a priori. It seems that there is no general natural metric for the usefulness of an instrument unlike other model selection problems such as , e.g., lag selection where such a metric is available.

The present paper provides a possible solution to the problem. If one views the set of 2^N possible sets of instruments as a space over which to minimise MSE for a given estimator then the problem becomes one of nonstandard minimisation of a function. The problem is nonstandard since the space over which minimisation occurs is discrete rather than continuous. A number of algorithms exist in the numerical analysis literature which suitably modified can be useful in this context. We focus on two distinct algorithms which provide a

theoretically valid and computationally tractable solution. One is simulated annealing and the second is genetic optimisation. Either of these two approaches can be used to minimise a function over a discrete domain and under suitable conditions is guaranteed to find the global minimum.

The paper is organised as follows: Section 2 sets out in detail the problem we would like to address. Section 3 presents details on the maximisation algorithms we consider. Section 4 presents a Monte Carlo exercise. Finally, Section 5 concludes.

2 Setup

This section presents the setup of the problem. The model considered is standard in the literature and given by

$$y_i = z_i'\alpha + x_{1,i}\beta + \epsilon_i \tag{1}$$

$$z_i = \Pi x_i + \eta_i \tag{2}$$

for $i = 1, \ldots, n$, where y_i is a scalar, z_i is a vector of variables possibly correlated with ϵ_i, x_i is a vector of exogenous variables uncorrelated with ϵ_i and η_i and $x_{1,i}$ is a d_1-dimensional subset of x_i. The aim is to estimate β. It is assumed that there exists a set of N instruments, denoted $\phi_i^N = (\phi_{1,i}, \ldots, \phi_{N,i})'$ which are (functions of) the x_i. We denote subsets of ϕ_i^N using binary notation. The reason for this will be made clear below. Thus, $\boldsymbol{\mathcal{J}}_j^N = \{\boldsymbol{\mathcal{J}}_{1,j}, \ldots, \boldsymbol{\mathcal{J}}_{N,j}\}$, where $\boldsymbol{\mathcal{J}}_{k,j} \in \{0, 1\}$, denotes the subset of instruments which contains the instruments $\phi_{k,i}$ for which $\boldsymbol{\mathcal{J}}_{k,j} = 1$. Of course, $j = 1, \ldots, 2^N$ and $\boldsymbol{\mathcal{J}}_j^N \in \{0, 1\}^N$. The vector of instruments contained in $\boldsymbol{\mathcal{J}}_j^N$ is denoted by $\phi_i^{\boldsymbol{\mathcal{J}}_j^N}$.

For a given subset of instruments, generically denoted $\boldsymbol{\mathcal{J}}$, we follow Donald and Newey (2001) and consider three well known IV estimators. To describe the estimators we define the following matrices: $\Phi^{\boldsymbol{\mathcal{J}}} = (\phi_1^{\boldsymbol{\mathcal{J}}}, \ldots, \phi_n^{\boldsymbol{\mathcal{J}}})'$, $P^{\boldsymbol{\mathcal{J}}} = \Phi^{\boldsymbol{\mathcal{J}}}(\Phi^{\boldsymbol{\mathcal{J}}\prime}\Phi^{\boldsymbol{\mathcal{J}}})^{-}\Phi^{\boldsymbol{\mathcal{J}}\prime}$, $y = (y_1, \ldots, y_n)'$, $Z = (z_1, \ldots, z_n)'$, $X_1 = (x_{1,1}, \ldots, x_{1,n})'$, $\delta = (\alpha', \beta')'$ and $W = (Y, X_1)$. A^- denotes an unspecified generalised inverse of A. Define also $\hat{\Lambda}$ to be the minimum of $(y - W\delta)'P^{\boldsymbol{\mathcal{J}}}(y - W\delta)/(y - W\delta)'(y - W\delta)$ and $\bar{\Lambda} = (\sum_{j=1}^{N} \boldsymbol{\mathcal{J}}_j - d_1 - 2)/n$. The three estimators considered are:

$$2SLS: \quad \hat{\delta} = (W'P^{\boldsymbol{\mathcal{J}}}W)^{-1}W'P^{\boldsymbol{\mathcal{J}}}y$$

$$LIML: \quad \hat{\delta} = (W'P^{\boldsymbol{\mathcal{J}}}W - \hat{\Lambda}W'W)^{-1}(W'P^{\boldsymbol{\mathcal{J}}}y - \hat{\Lambda}W'y)$$

$$B2SLS: \quad \hat{\delta} = (W'P^{\boldsymbol{\mathcal{J}}}W - \bar{\Lambda}W'W)^{-1}(W'P^{\boldsymbol{\mathcal{J}}}y - \bar{\Lambda}W'y)$$

where B2SLS denotes a bias adjusted version of 2SLS.

Under certain regularity conditions[1], Donald and Newey (2001), derive approximate estimators of the MSE of the three estimators of δ. For simplicity we will report these estimators for the case of a single right-hand side endogenous variable. These are given by:

$$2SLS: \quad \hat{S}(\boldsymbol{\mathcal{J}}) = \sigma^2_{1,\epsilon}\frac{(\sum_{j=1}^{N}\boldsymbol{\mathcal{J}}_j)^2}{n} + \sigma^2_{\epsilon}\left(\hat{R}(\boldsymbol{\mathcal{J}}) - \sigma^2_1\frac{\sum_{j=1}^{N}\boldsymbol{\mathcal{J}}_j}{n}\right)$$

$$LIML: \quad \hat{S}(\boldsymbol{\mathcal{J}}) = \sigma^2_{\epsilon}\left(\hat{R}(\boldsymbol{\mathcal{J}}) - \frac{\sigma^2_{1,\epsilon}}{\sigma^2_{\epsilon}}\frac{\sum_{j=1}^{N}\boldsymbol{\mathcal{J}}_j}{n}\right)$$

$$B2SLS: \quad \hat{S}(\boldsymbol{\mathcal{J}}) = \sigma^2_{\epsilon}\left(\hat{R}(\boldsymbol{\mathcal{J}}) + \frac{\sigma^2_{1,\epsilon}}{\sigma_{\epsilon}}\frac{\sum_{j=1}^{N}\boldsymbol{\mathcal{J}}_j}{n}\right)$$

where $\sigma^2_{\epsilon} = \tilde{\epsilon}'\tilde{\epsilon}/n$, $\sigma^2_1 = \tilde{u}_1'\tilde{u}_1/n$, $\sigma_{1,\epsilon} = \tilde{u}_1'\tilde{\epsilon}/n$, $\tilde{u}_1 = \tilde{u}\tilde{H}^{-1}$, $\tilde{H} = W'P^{\tilde{\boldsymbol{\mathcal{J}}}}W/n$, $\tilde{u} = (I - P^{\tilde{\boldsymbol{\mathcal{J}}}})W$, $\tilde{\epsilon} = y - W\tilde{\delta}$, $\hat{R} = \frac{1}{n}\sum_{i=1}^{n}\frac{\hat{u}_{1,i,}{}^2}{1-P^{\boldsymbol{\mathcal{J}}}_{ii}}$, $\hat{u} = (I - P^{\boldsymbol{\mathcal{J}}})W$, $\hat{u}_1 = \hat{u}\tilde{H}^{-1}$, $\boldsymbol{\mathcal{J}}$ denotes a generic instrument subset being evaluated and $\tilde{\boldsymbol{\mathcal{J}}}$ denotes some initially chosen instrument subset which is fixed across the minimisation of the MSE function over the instrument subsets. Given these MSE estimators Donald and Newey (2001) suggest that the appropriate number of instruments is chosen by minimising the estimated MSE. Monte Carlo evidence supports the suggested method.

The main problem with the minimisation of the estimated MSE concerns the choice of the search path over possible instrument subsets. As pointed out in the introduction, given a set of N instruments over which to choose a subset for inclusion in the estimation there exist 2^N possible subsets that can be considered. Inspecting all of them by estimating their asymptotic MSE is a computationally intractable task. To give an idea of the problem when $N = 50$ and optimistically assuming that 100000 instruments subsets can be evaluated per second, we still need about 357 years for an evaluation of all subsets. Furthermore, Donald and Newey (2001, pp. 1164) claim that not all subsets should be inspected as this will lead to a variable estimator of the chosen subset.

One solution is simply to use some economic theory to rank the instruments in order of relevance, sequentially augment the instrument set until all available instruments are consider and choose the subset that minimises estimated MSE. Symbolically, this is equivalent to searching over the following set of instrument subsets: $\left\{\boldsymbol{\mathcal{J}}^{N}_{1:j}\,|j = 1, \ldots, N\right\}$, where

[1]These conditions are Assumptions 1-3 of Donald and Newey (2001) plus restrictions on the rate of growth of N relative to n, satisfied if $N^2/n \to 0$.

$\mathcal{J}_{1:j}^{N} = \{\underbrace{1,\ldots,1}_{j},\underbrace{0,\ldots,0}_{N-j}\}$. This is analogous to the standard information criterion search for the appropriate lag order in time series analysis. However, such an approach is more questionable in the IV estimator context, since there is a natural ordering of lags to be included in a time series model which is lacking in the IV estimator context.

Perhaps, a more appropriate analogy is with the literature focussing on variable selection for regression models (see, e.g., Hoover and Perez (1999) and references cited therein). In the variable selection problem, economic theory provides some guidance on the choice of variables to be included in a regression. This guidance, however, is deemed inadequate to provide a full specification of regression models and therefore variable selection methods such as the widely used 'general-to-specific' approach, developed and popularised in a number of papers by David Hendry and his co-authors, such as Krolzig and Hendry (2001), have appeared in the literature. In fact, the problem in the IV estimator case is more acute. Whereas, in the variable selection problem economic theory is essentially asked to provide a list of variables whose coefficients are different from zero in a regression model, in the IV estimation problem the question is altogether more vague. Clearly, for an instrument to be useful its coefficient in the reduced form model (2) should be nonzero. But all valid instruments are assumed to have this property. Economic theory in this case, must have something to say on, e.g., the relative magnitude of such coefficients. Clearly such demands are unlikely to be met by current economic theory.

We suggest that the estimated approximate MSE functions of IV estimators be minimised in a similar way to other continuous objective functions such as the log-likelihood. Clearly, since the space, over which the function is to be minimised, is discrete, standard optimisation algorithms are not useful. However, there exist classes of algorithms, referred to as combinatorial algorithms, that can be used to minimise functions over discrete domains. The canonical domain for functions to be minimised using these algorithms is $\{0,1\}^N$. This makes clear the need for restating the problem in binary notation, as we did earlier in this section. These algorithms essentially provide a data dependent search path in $\{0,1\}^N$ which, under certain conditions, is guaranteed to contain the minimum without searching over all the elements of the domain.

3 Nonstandard Optimisation Algorithms

In the previous section we saw how the problem of choosing an instrument subset for IV estimation can be translated to a problem of minimising an estimate approximate MSE function. On the one hand the space where the MSE function is defined is discrete and hence standard optimisation methods cannot be applied. On the other hand, standard grid search which is usually implemented to minimise discrete functions, as in, e.g., lag selection, is clearly infeasible due to the computational burden of the problem. One alternative is to resort to nonstandard optimisation algorithms that do not require neither smoothness nor continuity for the algorithm to converge.

3.1 Simulated Annealing

Simulated annealing is a generic term used to refer to a family of powerful optimisation algorithms. In essence, it is a method that uses the objective function to create a non-homogeneous Markov chain that asymptotically converges to the optimum of the objective function. It is especially well suited for functions defined in discrete spaces like the MSE functions considered here. Below, we give a description of the algorithm together with the necessary arguments that illustrate its validity in our context. We describe the operation of the algorithm when the domain of the function (MSE function) is the set of binary strings i.e. $\{\boldsymbol{\mathcal{J}} = (\boldsymbol{\mathcal{J}}_1, \ldots, \boldsymbol{\mathcal{J}}_N)' | \boldsymbol{\mathcal{J}}_i \in \{0, 1\}\}$.

Each step of the algorithm works as follows starting from an initial string $\boldsymbol{\mathcal{J}}_0$.

1. Using $\boldsymbol{\mathcal{J}}_i$ choose a neighboring string at random, denoted $\boldsymbol{\mathcal{J}}^*_{i+1}$. We discuss the definition of a neighborhood below.

2. If $\hat{S}(\boldsymbol{\mathcal{J}}_i) > \hat{S}(\boldsymbol{\mathcal{J}}^*_{i+1})$, set $\boldsymbol{\mathcal{J}}_{i+1} = \boldsymbol{\mathcal{J}}^*_{i+1}$. Else, set $\boldsymbol{\mathcal{J}}_{i+1} = \boldsymbol{\mathcal{J}}^*_{i+1}$ with probability $e^{-(\hat{S}(\boldsymbol{\mathcal{J}}^*_{i+1}) - \hat{S}(\boldsymbol{\mathcal{J}}_i))/T_i}$ or set $\boldsymbol{\mathcal{J}}_{i+1} = \boldsymbol{\mathcal{J}}_i$ with probability $1 - e^{-(\hat{S}(\boldsymbol{\mathcal{J}}^*_{i+1}) - \hat{S}(\boldsymbol{\mathcal{J}}_i))/T_i}$.

Heuristically, the term T_i gets smaller making it more difficult, as the algorithm proceeds, to choose a point that does not decrease $\hat{S}(.)$. The issue of the neighborhood is extremely relevant. What is the neighborhood? Intuitively, the neighborhood could be the set of strings that differ from the current string by one element of the string. But this may be too restrictive. We can allow the algorithm to choose at random, up to some maximum integer (say h), the number of string elements at which the string at steps i and $i+1$ will differ. So the neighborhood is all strings with up to h different bits from the current string. Another issue is when to stop the algorithm. There are a number of alternatives in the literature. We have

chosen to stop the algorithm if it has not visited a string with lower $\hat{S}(.)$ than the current minimum for a prespecified number of steps (B_v) (Steps which stay at the same string do not count) or if the number of overall steps exceeds some other prespecified number (B_s). All strings visited by the algorithm are stored and the best chosen at the end rather than the final one.

The simulated annealing algorithm has been proven by Hajek (1988) (see also Del Moral and Miclo (1999)) to converge asymptotically, i.e. as $i \to \infty$, to the minimum of the function almost surely as long as $T_i = T_0/ln(i)$ for some T_0 for sufficiently large T_0. In particular, for almost sure convergence to the minimum it is required that $T_0 > d^*$. d^* denotes the maximum depth of all local minima of the function $\hat{S}(.)$. Heuristically, the depth of a local minimum, $\boldsymbol{\mathcal{J}}_1$, is defined as the smallest number $E > 0$, over all trajectories, such that the function never exceeds $\hat{S}(\boldsymbol{\mathcal{J}}_1) + E$ during a trajectory from[2] this minimum to any other local minimum, $\boldsymbol{\mathcal{J}}_2$, for which $\hat{S}(\boldsymbol{\mathcal{J}}_1) > \hat{S}(\boldsymbol{\mathcal{J}}_2)$.

3.2 The genetic algorithm (GA)

Once again, we describe the operation of the algorithm when the domain of the function is the set of binary strings. The motivating idea of genetic algorithms is to start with a population of binary strings which then evolve and recombine to produce new populations with 'better' characteristics, i.e. lower values for the MSE function. We start with an initial population represented by a $N \times m$ matrix made up of 0's and 1's. Columns represent strings. m is the chosen size of the population. Denote this population (matrix) by $\mathbf{P}_0$. The genetic algorithm involves defining a transition from $\mathbf{P}_i$ to $\mathbf{P}_{i+1}$. The algorithm has the following steps:

1. For $\mathbf{P}_i$ create a $m \times 1$ 'fitness' vector, $\mathbf{p}_i$, by calculating for each column of $\mathbf{P}_i$ its 'fitness'. The choice of the 'fitness' function is completely open and depends on the problem. For our purposes it is the opposite of the MSE function. Normalise $\mathbf{p}_i$, such that its elements lie in $(0, 1)$ and add up to 1. Denote this vector by $\mathbf{p}_i^*$. Treat $\mathbf{p}_i^*$ as a vector of probabilities and resample m times out of $\mathbf{P}_i$ with replacement, using the vector $\mathbf{p}_i^*$ as the probabilities with which each string with be sampled. So 'fit' strings are more likely to be chosen. Denote the resampled population matrix by $\mathbf{P}_{i+1}^1$.

[2]A trajectory from $\boldsymbol{\mathcal{J}}_1$ to $\boldsymbol{\mathcal{J}}_2$ is a set of strings, $\boldsymbol{\mathcal{J}}_{11}, \boldsymbol{\mathcal{J}}_{12}, \ldots, \boldsymbol{\mathcal{J}}_{1p}$, such that (i) $\boldsymbol{\mathcal{J}}_{11} \in N(\boldsymbol{\mathcal{J}}_1)$, (ii) $\boldsymbol{\mathcal{J}}_{1p} \in N(\boldsymbol{\mathcal{J}}_2)$ and (iii) $\boldsymbol{\mathcal{J}}_{1i+1} \in N(\boldsymbol{\mathcal{J}}_{1i})$ for all $i = 1, \ldots, p$, where $N(\boldsymbol{\mathcal{J}})$ denotes the set of strings that make up the neighborhood of $\boldsymbol{\mathcal{J}}$.

2. Perform cross over on $\mathbf{P}^1_{i+1}$. For cross over we do the following: Arrange all strings in $\mathbf{P}^1_{i+1}$, in pairs (assume that m is even). Denote a generic pair by $(a^\alpha_1, a^\alpha_2, \ldots, a^\alpha_n)$, $(a^\beta_1, a^\beta_2, \ldots, a^\beta_n)$. Choose a random integer between 2 and $n-1$. Denote this by j. Replace the pair by the following pair: $(a^\alpha_1, a^\alpha_2, \ldots, a^\alpha_j, a^\beta_{j+1}, \ldots, a^\beta_n)$, $(a^\beta_1, a^\beta_2, \ldots, a^\beta_j, a^\alpha_{j+1}, \ldots, a^\alpha_n)$. Perform cross over on each pair with probability p_c. Denote the new population by $\mathbf{P}^2_{i+1}$. Usually p_c is set to some number around 0.5-0.6.

3. Perform mutation on $\mathbf{P}^2_{i+1}$. This amounts to flipping the bits (0 or 1) of $\mathbf{P}^2_{i+1}$ with probability p_m. p_m is usually set to a small number, say 0.01. After mutation the resulting population is $\mathbf{P}_{i+1}$.

These steps are repeated a prespecified number of times (B_g). Each set of steps is referred to as generation in the genetic literature. If a string is to be chosen this is the one with maximum fitness. For every generation we store the identity of the string with maximum 'fitness'. At the end of the algorithm the string with the lowest MSE value over all members of the populations and all generations is chosen. One can think of the transition from one string of maximum fitness to another as a Markov Chain. So this is a Markov Chain algorithm. In fact, the Markov chain defined over all possible strings is time invariant but not irreducible as at least the $m-1$ least fit strings will never be picked. To see this note that in any population there will be a string with more fitness than that of the $m-1$ worst strings. There has been considerable work on the theoretical properties of genetic algorithms. Hartl and Belew (1990) and Del Moral and Miclo (1999) have shown that with probability approaching one, the population at the B-th generation will contain the global maximum as $B \to \infty$. For more details see also Del Moral, Kallel, and Rowe (2001).

4 Monte Carlo Study

4.1 Monte Carlo Setup

In order to illustrate the potential of the new methods we carry out a Monte Carlo study. The study follows elements of the setup of Donald and Newey (2001) for comparability purposes. The model is given by

$$y_i = \alpha z_i + \epsilon_i \tag{3}$$

$$z_i = x_i'\pi + \eta_i \tag{4}$$

We set $\alpha = 0.1$ and consider $n = 100, 500$. We also set $N = 20$. Important parameters for the performance of the estimators are the covariance of ϵ_i and η_i denoted $\sigma_{\eta,\epsilon}$ which is set to

$\sigma_{\eta,\epsilon} \in \{0.1, 0.5, 0.9\}$ and the R^2 of model (4) which we set to 0.01 or 0.1. Following Donald and Newey (2001), one way for setting R^2 in (4) is to use the formula

$$\pi_k = \sqrt{\frac{R^2}{N(1-R^2)}}, k = 1, \ldots, N$$

where $\pi = (\pi_1, \ldots, \pi_N)'$.

In this Monte Carlo study we concentrate on the simulated annealing algorithm. The reasons for this are computational tractability and prior experience with the two algorithms. Kapetanios (2004b) and Kapetanios (2004a) analyse the performance of the algorithms in other problems in econometrics[3]. In both cases it is found that simulated annealing outperforms the genetic algorithm in terms of finding the optimum for the function being optimised. Further, reasonable choices for the parameters of the search algorithms indicate that simulated annealing maybe computable cheaper than the genetic algorithm by a factor of about 10. To see this note that setting $B_s = 2000$, $B_v = 500$ and $h = 1$ for the simulated annealing algorithm (which are the choices of parameters we make) leads to 2000 separate IV estimations. By comparison setting $m = 200$ and $B_g = 100$ for the genetic algorithm, which are reasonable choices relative to the literature we have 20000 IV estimations.

Note that although 2000 evaluations may appear high for the n, N combinations we consider, there is no need for B_g to grow with N as prior experience suggests limited sensitivity of the algorithm to B_g as long as it is reasonably large to begin with. We carry out 1000 Monte Carlo replications. Again anything significantly more than that is prohibitively expensive. Note that the results reported here took more than 1/2 months of computer time on a personal computer with 3 Ghz processor speed.

4.2 Results

We present the following statistics on the estimated coefficients: Firstly, we present the mean square error of the estimators over the replications. Secondly we present the median bias, thirdly we present the median absolute deviation from the true value and finally the range between the 90% and 10% quantiles. We choose to present results on MSE, unlike Donald and Newey (2001), as it seems to be the natural choice for a reporting medium on the performance of a method designed to minimise MSE. We consider the estimators B2SLS (denoted BSLS in the Tables), 2SLS and LIML and three ways of choosing the instruments.

[3]Kapetanios (2004b) looks at variable selection in regression models whereas Kapetanios (2004a) looks at cluster analysis for panel datasets.

The first includes all available instruments and is denoted by the subscript a, the second chooses the order of instruments according to Donald and Newey (2001) (denoted by the subscript o) and the third uses simulated annealing to minimise MSE and is denoted by the subscript s. Results are presented in Tables 1 and 2.

Table 1 presents results for $n = 100$. Results, as expected, have a tendency to get worse for all estimators when R^2 falls and $\sigma_{\eta,\epsilon}$ rises in absolute value. Looking at B2SLS first we note that $BSLS_o$ seems to improve on $BSLS_a$ overall with bias results being more mixed than for other performance indicators. $BSLS_s$ provides clear further improvement on $BSLS_o$ for most cases and most indicators. Moving on the 2SLS we see that $2SLS_o$ is usually doing worse than $BSLS_a$ and in some cases much worse. $2SLS_s$ improves greatly on $2SLS_o$ but not on $BSLS_a$ where the comparison is more mixed. Looking at LIML we see that $LIML_a$ is comparable to $LIML_o$ with $LIML_s$ performing better. Clearly, minimising MSE using a optimisation algorithm is helpful for the performance of all estimators at this sample size.

Table 2 looks at the performance of the estimators for a sample size of $n = 500$. Clearly all estimators do better for this sample size as expected. Again $BSLS_o$ improves drastically on $BSLS_a$ in a majority of cases with $BSLS_s$ providing further improvement. For $2SLS$ instrument selection does not appear to be that helpful with $2SLS_s$ dominating $BSLS_o$ in most cases. Finally, for LIML $LIML_s$ improves greatly upon both $LIML_o$ and $LIML_a$. Overall, a clear conclusion emerges for the superiority of selecting instruments by minimising MSE via simulated annealing.

5 Conclusion

Estimation by Instrumental Variables is extremely common in the econometric literature. A major preoccupation concerns the choice of the instruments used in the estimation. This choice has two related components. Firstly, one must choose a set of valid instruments among all possible instruments and secondly one must choose among all valid instruments so as to optimise the performance of a given estimator. This paper addresses the second component.

Clearly, an obvious method for guiding the selection involves using economic theory. Nevertheless this is likely to be of little help in general. Donald and Newey (2001) has suggested a method for choosing the number of instruments used so as to minimise the MSE of the estimator. Nevertheless, the ordering of the instruments used in choosing the number of

instruments to include is an issue. We suggest using nonstandard optimisation algorithms to optimise the search for the subset of instruments that minimises the MSE of a given IV estimator.

After discussing the optimisation algorithms we suggest, we present a Monte Carlo study similar to that in Donald and Newey (2001) which illustrates the potential of the new methods. Further research should concentrate on an empirical evaluation of the new methods as well as exploring their potential when applied to other IV estimators.

References

Andrews, D. W. K. (1999): "Consistent Moment Selection Procedures for Generalised Method of Moments," *Econometrica*, 67, 543–564.

Bound, J., D. Jaeger, and R. Baker (1996): "Problems with Instrumental Variables Estimation When the Correlation Between Instruments and the Endogenous Explanatory Variable is Weak," *Journal of the American Statistical Association*, 90, 443–450.

Del Moral, P., L. Kallel, and J. Rowe (2001): "Modelling Genetic Algorithms with Interacting Partical Systems," *Revista de Matematica, Teoria y Aplicaciones*, 8(2).

Del Moral, P., and L. Miclo (1999): "On the Convergence and the Applications of Generalised Simulated Annealing," *SIAM Journal of Control and Optimisation*, 37(4), 1222–1250.

Donald, S. G., and W. K. Newey (2001): "Choosing the Number of Instruments," *Econometrica*, 69(5), 1161–1191.

Hajek, B. (1988): "Cooling Schedules for Optimal Annealing," *Mathematics of Operations Research*, 13(2), 311–331.

Hartl, H. R. F., and R. K. Belew (1990): "A Global Convergence Proof for a Class of Genentic Algorithms," *Technical Report, Technical University of Vienna*.

Hoover, K. D., and S. J. Perez (1999): "Data Mining Reconsidered: Encompassing and the General-To-Specific Approach to Specification Search," *Econometrics Journal*, 2, 167–191.

Kapetanios, G. (2004a): "Cluster Analysis of Panel Datasets using Non-Standard Optimisation of Information Criteria," *Mimeo, Queen Mary, University of London*.

——— (2004b): "Variable Selection using Non-Standard Optimisation of Information Criteria," *Mimeo, Queen Mary, University of London*.

Krolzig, H. M., and D. F. Hendry (2001): "Computer Automation of General-to-Specific Model Selection Procedures," *Journal of Economic Dynamics and Control*, 25(6–7), 831–866.

Morimune, K. (1983): "Approximate Distributions of k-Class Estimators When the Degree of Overidentifiability is Large Compared with the Sample Size," *Econometrica*, 51, 821–841.

Nagar, A. L. (1959): "The Bias and Moment Matrix of the General k-Class Estimators of the Parameters in Simultaneous Equations," *Econometrica*, 27, 575–595.

Rothenberg, T. J. (1983): "Asymptotic Properties of Some Estimators in Structural Models," in *Studies in Econometrics, Time Series and Multivariate Statistics*, ed. by S. Karlin, T. Amemiya, and L. A. Goodman. Academic Press.

Table 1: $n = 100$									
		MSE	$\overset{Med.}{Bias}$	$\overset{Med.}{AD}$	$\overset{Dec.}{Rge}$	MSE	$\overset{Med.}{Bias}$	$\overset{Med.}{AD}$	$\overset{Dec.}{Rge}$
		$R^2 = 0.1$				$R^2 = 0.01$			
0.1	$BSLS_a$	1.277	0.016	0.328	1.523	3.388	0.124	0.583	3.225
	$BSLS_o$	1.135	0.041	0.312	1.481	2.376	0.098	0.520	2.447
	$BSLS_s$	1.109	0.027	0.311	1.271	2.203	0.119	0.460	2.082
	$2SLS_a$	0.040	0.061	0.132	0.477	0.065	0.110	0.165	0.571
	$2SLS_o$	0.205	0.077	0.174	0.688	1.723	0.088	0.503	2.265
	$2SLS_s$	0.078	0.066	0.168	0.636	0.889	0.128	0.374	1.550
	$LIML_a$	1.669	0.006	0.352	1.713	5.421	0.133	0.778	4.271
	$LIML_o$	3.819	0.082	0.658	3.510	5.955	0.140	0.861	4.728
	$LIML_s$	1.393	0.024	0.319	1.423	3.352	0.119	0.544	2.900
0.5	$BSLS_a$	1.449	0.137	0.348	1.515	4.171	0.493	0.768	3.400
	$BSLS_o$	1.302	0.142	0.346	1.431	2.232	0.437	0.650	2.490
	$BSLS_s$	1.047	0.180	0.326	1.197	1.931	0.424	0.583	1.972
	$2SLS_a$	0.133	0.324	0.324	0.428	0.256	0.474	0.474	0.531
	$2SLS_o$	0.615	0.315	0.377	1.104	1.618	0.487	0.631	2.051
	$2SLS_s$	0.163	0.301	0.315	0.664	0.886	0.454	0.528	1.428
	$LIML_a$	1.091	0.058	0.315	1.496	4.233	0.307	0.823	3.874
	$LIML_o$	3.979	0.271	0.758	3.482	5.641	0.424	0.923	4.344
	$LIML_s$	1.115	0.117	0.317	1.283	2.849	0.330	0.628	2.617
0.9	$BSLS_a$	2.073	0.188	0.360	1.587	2.239	0.814	0.864	1.776
	$BSLS_o$	1.785	0.220	0.360	1.440	1.687	0.821	0.847	1.253
	$BSLS_s$	0.927	0.279	0.359	1.064	1.879	0.775	0.796	1.070
	$2SLS_a$	0.346	0.580	0.580	0.287	0.743	0.850	0.850	0.281
	$2SLS_o$	0.831	0.521	0.570	1.284	1.358	0.842	0.851	1.159
	$2SLS_s$	0.359	0.516	0.522	0.568	0.817	0.798	0.804	0.711
	$LIML_a$	0.797	0.000	0.244	1.137	3.542	0.623	0.767	2.664
	$LIML_o$	3.316	0.472	0.688	3.183	3.124	0.809	0.916	2.340
	$LIML_s$	0.583	0.154	0.259	0.929	2.021	0.690	0.747	1.572

Table 1: $n = 500$									
		MSE	$\overset{Med.}{Bias}$	$\overset{Med.}{AD}$	$\overset{Dec.}{Rge}$	MSE	$\overset{Med.}{Bias}$	$\overset{Med.}{AD}$	$\overset{Dec.}{Rge}$
		$R^2 = 0.1$				$R^2 = 0.01$			
0.1	$BSLS_a$	0.029	-0.006	0.100	0.407	2.721	0.019	0.499	2.672
	$BSLS_o$	0.027	-0.003	0.104	0.406	1.160	0.097	0.390	1.729
	$BSLS_s$	0.024	0.014	0.102	0.378	0.102	0.081	0.200	0.771
	$2SLS_a$	0.014	0.022	0.077	0.295	0.052	0.054	0.148	0.549
	$2SLS_o$	0.062	0.037	0.096	0.401	1.423	0.089	0.416	2.083
	$2SLS_s$	0.020	0.032	0.091	0.334	0.115	0.074	0.200	0.783
	$LIML_a$	0.030	-0.008	0.105	0.415	3.035	-0.049	0.538	2.787
	$LIML_o$	1.930	0.019	0.371	1.881	5.318	0.085	0.764	4.188
	$LIML_s$	0.025	0.009	0.104	0.395	0.277	0.078	0.232	0.883
0.5	$BSLS_a$	0.031	0.010	0.103	0.400	2.610	0.242	0.468	2.541
	$BSLS_o$	0.027	0.036	0.104	0.398	0.906	0.389	0.466	1.442
	$BSLS_s$	0.029	0.107	0.124	0.340	0.218	0.378	0.379	0.663
	$2SLS_a$	0.028	0.132	0.135	0.273	0.184	0.393	0.393	0.449
	$2SLS_o$	0.199	0.165	0.221	0.651	1.311	0.425	0.556	1.806
	$2SLS_s$	0.055	0.202	0.204	0.327	0.280	0.391	0.395	0.691
	$LIML_a$	0.026	0.005	0.098	0.389	3.343	0.095	0.453	2.513
	$LIML_o$	2.154	0.045	0.428	2.031	5.058	0.354	0.839	4.096
	$LIML_s$	0.028	0.093	0.118	0.352	0.304	0.350	0.363	0.734
0.9	$BSLS_a$	0.033	0.026	0.110	0.425	2.803	0.454	0.581	2.418
	$BSLS_o$	0.027	0.064	0.119	0.391	0.828	0.660	0.669	0.980
	$BSLS_s$	0.045	0.191	0.191	0.277	0.509	0.683	0.683	0.411
	$2SLS_a$	0.060	0.233	0.233	0.223	0.529	0.718	0.718	0.288
	$2SLS_o$	0.325	0.240	0.312	0.902	1.071	0.706	0.727	1.297
	$2SLS_s$	0.145	0.373	0.373	0.257	0.560	0.718	0.718	0.439
	$LIML_a$	0.022	0.004	0.095	0.365	2.537	0.092	0.367	1.873
	$LIML_o$	2.410	0.055	0.401	1.987	3.313	0.679	0.833	2.876
	$LIML_s$	0.033	0.152	0.153	0.261	0.671	0.607	0.610	0.522

This working paper has been produced by
the Department of Economics at
Queen Mary, University of London

Department of Economics
Queen Mary, University of London
Mile End Road
London E1 4NS
Tel: +44 (0)20 7882 5096
Fax: +44 (0)20 8983 3580
Web: www.econ.qmul.ac.uk/papers/wp.htm