Econometrics Journal (1999), volume **2**, pp. 167–191.

Data mining reconsidered: encompassing and the general-to-specific approach to specification search

KEVIN D. HOOVER, STEPHEN J. PEREZ

Department of Economics, University of California, Davis, California 95616-8578, USA
E-mail: `kdhoover@ucdavis.edu`; Homepage: `www.ucdavis.edu/~kdhoover/`
Department of Economics, Washington State University, Pullman, Washington 99164-4741, USA
E-mail: `sjperez@wsu.edu`; Homepage: `www.cbe.wsu.edu/~sjperez/`

Summary This paper examines the efficacy of the general-to-specific modeling approach associated with the LSE school of econometrics using a simulation framework. A mechanical algorithm is developed which mimics some aspects of the search procedures used by LSE practitioners. The algorithm is tested using 1000 replications of each of nine regression models and a data set patterned after Lovell's (1983) study of data mining. The algorithm is assessed for its ability to recover the data-generating process. Monte Carlo estimates of the size and power of exclusion tests based on t-statistics for individual variables in the specification are also provided. The roles of alternative sizes for specification tests in the algorithm, the consequences of different signal-to-noise ratios, and strategies for reducing overparameterization are also investigated. The results are largely favorable to the general-to-specific approach. In particular, the size of exclusion tests remains close to the nominal size used in the algorithm despite extensive search.

Keywords: *General-to-specific, Encompassing, Data mining, LSE econometrics.*

1. INTRODUCTION

In recent years a variety of competing econometric methodologies have been debated: among others, structural modeling, vector autoregressions, calibration, extreme-bounds analysis, and the so-called LSE [London School of Economics] approach.[1] In this study, we evaluate the last of these, the LSE approach—not philosophically, theoretically or methodologically, but practically. We pose the question: in a simulation study in which we know the underlying process that generated the data, do the methods advocated by David Hendry and other practitioners of the LSE econometric methodology in fact recover the true specification?[2] A doubt often felt, and sometimes articulated, about the LSE approach is that it amounts to systematized 'data mining'. The practice of data mining has itself been scrutinized only infrequently (e.g. Mayer (1980,

[1] See Ingram (1995), Canova (1995), Mizon (1995), Kydland and Prescott (1995), and Leamer (1983) for overviews.

[2] The adjective 'LSE' is, to some extent, a misnomer. It derives from the fact that there is a tradition of time-series econometrics that began in the 1960s at the London School of Economics; see Mizon (1995) for a brief history. The practitioners of LSE econometrics are now widely dispersed among academic institutions throughout Britain and the world.

1993), Cox (1982), Leamer (1983, 1985), Lovell (1983), Chatfield (1995), Hoover (1995), Nester (1996)).

Lovell (1983) makes one of the few attempts that we know of to evaluate specification search in a simulation framework. Unfortunately, none of the search algorithms that he investigates comes close to approximating LSE methodology. Still, Lovell's simulation framework provides a neutral test-bed on which we evaluate LSE methods, one in which there is no question of our having 'cooked the books'. Within this framework, we pose a straightforward question: does the LSE approach work?

2. ENCOMPASSING AND THE PROBLEM OF DATA MINING

The relevant LSE methodology is the *general-to-specific* modeling approach.[3] It relies on an intuitively appealing idea. A sufficiently complicated model can, in principle, describe the salient features of the economic world.[4] Any more parsimonious model is an improvement on such a complicated model if it conveys *all* of the same information in a simpler, more compact form. Such a parsimonious model would necessarily be superior to all other models that are restrictions of the completely general model except, perhaps, to a class of models nested within the parsimonious model itself. The art of model specification in the LSE framework is to seek out models that are valid parsimonious restrictions of the completely general model, and that are not redundant in the sense of having an even more parsimonious models nested within them that are also valid restrictions of the completely general model.

The name 'general-to-specific' itself implies the contrasting methodology. The LSE school stigmatizes much of common econometric practice as *specific-to-general*. Here one starts with a simple model, perhaps derived from a simplified (or highly restricted) theory. If one finds econometric problems (e.g. serial correlation in the estimated errors) then one complicates the model in a manner intended to solve the problem at hand (e.g. one postulates that the error follows a first-order autoregressive process (AR(1)) of a particular form, so that estimation using a Cochrane–Orcutt procedure makes sense).

The general-to-specific modeling approach is related to the theory of *encompassing*.[5] Roughly speaking, one model encompasses another if it conveys all of the information conveyed by another model. It is easy to understand the fundamental idea by considering two non-nested models of the same dependent variable. Which is better? Consider a more general model that uses the non-redundant union of the regressors of the two models. If model I is a valid restriction of the more general model (e.g. based on an F-test), and model II is not, then model I encompasses model II. If model II is a valid restriction and model I is not, then model II encompasses model I. In either case, we know everything about the joint model from one of the restricted models; we therefore know everything about the other restricted model from the one. There is, of course, no necessity that either model will be a valid restriction of the joint model: each could convey

[3]The LSE approach is described sympathetically in Gilbert (1986), Hendry (1995, 1997, esp. Chs 9–15), Pagan (1987), Phillips (1988), Ericsson *et al.* (1990), and Mizon (1995). For more sceptical accounts, see Hansen (1996) and Faust and Whiteman (1995, 1997) to which Hendry (1997) replies.

[4]This is a truism. Practically, however, it involves a leap of faith; for models that are one-to-one, or even distantly approach one-to-one, with the world are not tractable.

[5]For general discussions of encompassing, see, for example, Mizon (1984, 1995), Hendry and Richard (1987) and Hendry (1988, 1995, Ch. 14).

information that the other failed to convey. In population, a necessary, but not sufficient, condition for one model to encompass another is that it have a lower standard error of regression.[6]

A hierarchy of encompassing models arises naturally in a general-to-specific modeling exercise. A model is tentatively admissible on the LSE view if it is congruent with the data in the sense of being: (i) consistent with the measuring system (e.g. not permitting negative fitted values in cases in which the data are intrinsically positive), (ii) coherent with the data in that its errors are innovations that are white noise as well as a martingale difference sequence relative to the data considered, and (iii) stable (cf. Phillips (1988, pp. 352–353), White (1990, pp. 370–374), Mizon (1995, pp. 115–122)). Further conditions (e.g. consistency with economic theory, weak exogeneity of the regressors with respect to parameters of interest, orthogonality of decision variables) may also be required for economic interpretability or to support policy interventions or other particular purposes. While consistency with economic theory and weak exogeneity are important components of the LSE methodology, they are not the focus here and are presumed in the simulation study. If a researcher begins with a tentatively admissible general model and pursues a chain of simplifications, at each step maintaining admissibility and checking whether the simplified model is a valid restriction of the more general model, then the simplified model will be a more parsimonious representation of all the models higher on that particular chain of simplification and will encompass all of the models lower along the same chain.

The first charge against the general-to-specific approach as an example of invidious data mining points out that the encompassing relationships that arise so naturally apply only to a specific path of simplifications. There is no automatic encompassing relationship between the final models of different researchers who have wandered down different paths in the forest of models nested in the general model. One answer to this is that any two models can be tested for encompassing, either through the application of non-nested hypothesis tests or through the approach described above of nesting them within a joint model. Thus, the question of which, if either, encompasses the other can always be resolved. Nevertheless, critics may object—with some justification—that such playoffs are rare and do not consider the entire range of possible termini of general-to-specific specification searches. We believe that this is an important criticism and we will return to it presently.

A second objection notes that variables may be correlated either because there is a genuine relation between them or because—in short samples—they are adventitiously correlated. Thus, a methodology that emphasizes choice among a wide array of variables based on their correlations is bound to select variables that just happen to be related to the dependent variable in the particular data set, even though there is no economic basis for the relationship. This is the objection of Hess *et al*. (1998) that the general-to-specific specification search of Baba *et al.* (1992) selects an 'overfitting' model.

By far the most common reaction of critical commentators and referees to the general-to-specific approach questions the meaning of the test statistics associated with the final model. The implicit argument runs something like this: conventional test statistics are based on independent draws. The sequence of tests (F- or t-tests) on the same data used to guide the simplification of the general model, as well as the myriad of specification tests used repeatedly to check tentative admissibility, are necessarily not independent. The test statistics for any specification that has survived such a process are necessarily going to be 'significant'. They are 'Darwinian' in the sense that only the fittest survive. Since we know in advance that they pass the tests, the critical

[6] Economists, of course, do not work with populations but samples, often relatively small ones. Issues about the choice of the size of the tests and related matters are as always of great practical importance.

values for the tests could not possibly be correct. The critical values for such Darwinian test statistics must in fact be much higher, but just how much higher no one can say.

The LSE approach takes a different view of data mining. The difference can be understood by reflecting on a theorem proved by White (1990, pp. 379–380). The upshot of White's theorem is this: for a fixed set of specifications and a battery of specification tests, as the sample size grows toward infinity and increasingly smaller test sizes are employed, the test battery will—with a probability approaching unity—select the correct specification from the set. In such cases, White's theorem implies that type I and type II errors both fall asymptotically to zero. White's theorem states that, given enough data, only the true specification will survive a stringent enough set of tests. Another way to think about this is to say that a set of tests and a set of sample information restricts the class of admissible models. As we obtain more information, then this class can be further and further restricted; fewer and fewer models survive. This then turns the criticism of Darwinian test statistics on its head. The critics fear that the survivor of sequential tests survives accidentally and, therefore, that the critical values of such tests ought to be adjusted to reflect the likelihood of an accident. White's theorem suggests that the true specification survives precisely because the true specification is necessarily, in the long run, the fittest specification. Of course, White's theorem is an asymptotic result. It supports the general-to-specific approach in that it provides a vision of the idea of the true model as the one that is robust to increasing information. However, because it is an asymptotic result, it is not enough to assure us that LSE methods generate good results in the size of samples with which economists typically work. To investigate its practical properties we use Lovell's simulation framework.

3. THE 'MINE': LOVELL'S FRAMEWORK FOR THE EVALUATION OF DATA MINING

To investigate data mining in a realistic context Lovell (1983) begins with 20 annual macroeconomic variables covering various measures of real activity, government fiscal flows, monetary aggregates, financial market yields, labor market conditions and a time trend. These variables form the 'data mine', the universe for the specification searches that Lovell conducts. The advantage of such a data set is that it presents the sort of naturally occurring correlations (true and adventitious, between different variables and between the same variables through time) that practicing macroeconomists in fact face.

The test-bed for alternative methods of specification search is nine econometric models. The dependent variable for each specification is a 'consumption' variable artificially generated from a subset of between zero and two of the variables from the set of 20 variables plus a random error term. The random error term may be either independently normally distributed or autoregressive of order one. Except for one specification in which the dependent variable is purely random, the coefficients of Lovell's models were initially generated by regressing actual consumption on the various subsets of dependent variables or as linear combinations of models so generated. These subsets emphasize either monetary variables or fiscal variables. These coefficients are then used, together with a random number generator, to generate simulated dependent variables.[7]

[7] This was an attempt to add a bit of realism to the exercise by echoing the debate in the 1960s between Milton Friedman and David Meiselman, who stressed the relative importance of monetary factors in the economy, and the Keynesians, who stressed fiscal factors. While this is no longer a cutting-edge debate in macroeconomics, that in no way diminishes the usefulness of Lovell's approach as a method of evaluating specification search techniques.

For each of the nine specifications, Lovell created 50 separate artificial dependent 'consumption' variables corresponding to 50 independent draws for the random error terms. For each of these replications he then compared the ability to recover the true specification of three algorithms searching over the set of 20 variables. The three algorithms were stepwise regression, maximum $\bar{R}^2$, and choosing the subset of variables for which the minimum t-statistic of the subset is maximized relative to the minimums of the other subsets.

Lovell presents detailed analyses of the relative success of the different algorithms. He concludes that the results were not in general favorable to the success of data mining. With a nominal test size of 5%, the best of the three algorithms, step-wise regression, chose the correct variables only 70% of the time and was subject to a 30% rate of type I error.

To evaluate the general-to-specific approach, we modify Lovell's framework in three respects. First, we update his data to 1995. Using annual observations, as Lovell does, we repeated his simulations and found closely similar results on the new data set. Second, we substituted quarterly for annual data for each series to render the data similar to the most commonly used macroeconomic time-series. Again, we repeated Lovell's simulations on quarterly data and found results broadly similar to his. Finally, it has become more widely appreciated since Lovell's paper that numerous econometric problems arise from failing to account for non-stationarity in time-series data.[8] To avoid the issues associated with non-stationarity and cointegration, we differenced each series as many times as necessary to render it stationary (judged by Phillips and Perron's 1988 test).

Table 3 presents nine models constructed in the same manner as Lovell's but using the new stationary, quarterly data set.[9] Model 1 is purely random. Model 3 takes the log of simulated consumption as the dependent variable and is an AR(2) time-series model. Model 4 relates consumption to the M1 monetary aggregate, model 5 to government purchases, and model 6 to both M1 and government purchases. The dynamic models 2, 7, 8, and 9 are the same as the static models 1, 4, 5, and 6 except that an AR(1) error term replaces the identically, independently normally distributed error term. The principal question of this paper is, how well does the general-to-specific approach do at recovering these nine models in the universe of variables described in Table 1?

The universe of data for the evaluation of the general-to-specific approach is reported in Table 1. Notice that there are now only 18 primary variables reported: the time trend (one of Lovell's variables) is no longer relevant because the data are constructed to be stationary; furthermore, because of limitations in the sources of data, we omit Lovell's variable 'potential level of GNP in \$1958'.[10] Corresponding to each of the variables 1–18 are their lagged values numbered 19–36. In addition, variables 37–40 are the first to fourth lags of the 'consumption' variable.[11] Table 2 is the correlation matrix for variables 1–18 plus actual personal consumption expenditure.

[8] For surveys of non-stationary econometrics, see Stock and Watson (1988), Dolado *et al* (1990), Campbell and Perron (1991), and Banerjee (1995).

[9] All simulations are conducted using Matlab (version 5.1) and its normal random number generator.

[10] We also replaced Lovell's variables 'index, five coincident indicators' with 'index, four coincident indicators' and 'expected investment expenditure' with 'gross private investment'.

[11] As lags of the artificially generated dependent variables, these variables differ from model to model in the simulations below. Actual personal consumption expenditure is used in calibrating the models in Table 3.

Table 1. Candidates variables for specification search.

Variable	Variable number					Times differenced for stationarity[a]	CITIBASE identifier[b]
	Current	Lag					
		1	2	3	4		
Index of four coincident indicators	1	19				1	DCOINC
GNP price deflator	2	20				2	GD
Government purchases of goods and services	3	21				2	GGEQ
Federal purchases of goods and services	4	22				1	GGFEQ
Federal government receipts	5	23				2	GGFR
GNP	6	24				1	GNPQ
Disposable personal income	7	25				1	GYDQ
Gross private domestic investment	8	26				1	GPIQ
Total member bank reserves	9	27				2	FMRRA
Monetary base (federal reserve bank of St. Louis)	10	28				2	FMBASE
M1	11	29				1	FM1DQ
M2	12	30				1	FM2DQ
Dow Jones stock price	13	31				1	FSDJ
Moody's AAA corporate bond yield	14	32				1	FYAAAC
Labor force (16 years+, civilian)	15	33				1	LHC
Unemployment rate	16	34				1	LHUR
Unfilled orders (manufacturing, all industries)	17	35				1	MU
New orders (manufacturing, all industries)	18	36				2	MO
Personal consumption expenditure[c]	N/A	37	38	39	40	1	GCQ

Note: Data run 1959.1–1995.1. All data from CITIBASE: Citibank economic database (Floppy disk version), July 1995 release. All data converted to quarterly by averaging or summing as appropriate. All dollar denominated data in billions of constant 1987 dollars. Series FMRRA, FMBASE, GGFR, FSDJ, MU, and MO are deflated using the GNP price deflator (Series GD). [a] Indicates the number of times the series had to be differenced before a Phillips–Perron test could reject the null hypothesis of non-stationarity at a 5% significance level (Phillips and Perron 1988). [b] Indicates the identifier code for this series in the CITIBASE economic database. [c] For calibrating models in Table 4 actual personal consumption expenditure data is used as the dependent variables; for specification searches, actual data is replaced by artificial data generating according to models in Table 3. Variable numbers refer to these artificial data, which vary from context to context.

Table 2. Correlation matrix for search variables.

Variable name and number	Variable number																		
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	Dep.*
1. Four coincident indicators	**0.67**																		
2. GNP price deflator	*0.21*	***0.24***																	
3. Government purchases of goods and services	0.04	−0.09	**8.81**																
4. Federal purchases of goods and services	*−0.07*	*−0.08*	*0.54*	***6.22***															
5. Federal government receipts	0.21	0.28	0.03	0.01	**22.16**														
6. GNP	*0.83*	*0.16*	*0.13*	*0.03*	*0.20*	***30.71***													
7. Disposable personal income	0.57	0.07	0.07	−0.09	0.06	0.49	**25.09**												
8. Gross private domestic investment	*0.76*	*0.19*	*0.03*	*−0.18*	*0.13*	*0.83*	*0.40*	***25.91***											
9. Total member bank reserves	−0.02	0.24	0.07	0.14	0.40	−0.03	0.24	−0.16	**514.26**										
10. Monetary base (federal reserve bank of St. Louis)	*−0.02*	*0.49*	*−0.02*	*0.07*	*0.25*	*−0.06*	*0.10*	*−0.06*	*0.54*	***1.38***									
11. M1	0.24	−0.04	−0.04	0.00	0.16	0.27	0.17	0.17	0.25	0.20	**8.49**								
12. M2	*0.20*	*−0.06*	*−0.08*	*0.07*	*0.11*	*0.20*	*0.17*	*0.08*	*0.21*	*0.14*	*0.60*	***25.08***							
13. Dow Jones stock price	−0.04	−0.06	−0.06	−0.06	−0.12	0.03	−0.03	−0.02	−0.08	0.01	0.27	0.04	**95.40**						
14. Moody's AAA corporate bond yield	*0.23*	*0.11*	*−0.04*	*−0.05*	*0.07*	*0.11*	*0.07*	*0.20*	*−0.16*	*−0.06*	*−0.33*	*−0.33*	*−0.26*	***0.42***					
15. Labor force (16 years+, civilian)	0.17	0.04	0.03	−0.04	−0.03	0.11	0.09	0.07	−0.17	0.01	−0.04	−0.07	0.13	0.11	**321.15**				
16. *Unemployment rate*	*−0.85*	*−0.13*	*−0.01*	*−0.02*	*−0.09*	*−0.73*	*−0.31*	*−0.66*	*0.08*	*0.07*	*−0.23*	*−0.22*	*0.02*	*−0.22*	*0.02*	***0.35***			
17. Unfilled orders (manufacturing, all industries)	0.21	0.24	−0.08	0.04	0.03	0.16	0.05	0.10	−0.10	0.09	−0.39	−0.21	0.06	0.27	0.14	−0.23	**6248.9**		
18. New orders (manufacturing, all industries)	*0.23*	*0.12*	*−0.29*	*−0.15*	*0.25*	*0.22*	*0.15*	*0.10*	*0.21*	*0.01*	*0.28*	*0.19*	*0.06*	*0.12*	*0.01*	*−0.12*	*−0.04*	***4114.8***	
*Dep. personal consumption expenditure	0.60	−0.02	−0.02	−0.02	0.15	0.65	0.40	0.30	0.07	−0.03	0.47	0.41	0.18	−0.05	0.13	−0.50	−0.01	0.39	***15.85***

Note: Variables are differenced as indicated in Table 1. Elements in bold type on the main diagonals are the standard deviations of each variable for the period beginning 1959.2 or 1959.3, depending on the number of differences. Off-diagonal elements correlations are calculated for the variables in Table 1 for the period 1959.3 to 1995.1. *Dep. indicates that personal consumption expenditure is the dependent variable used in calibrating the models in Table 3. It is not a search variable. The dependent variables and its lags used in the simulations below are constructed according to those models.

Table 3. Models used to generate alternative artificial consumption-dependent variables.

Random errors		
$u_t \sim N(0, 1)$		
$u_t^* = 0.75u_{t-1}^* + u_t\sqrt{7/4}$		
Models		
Model 1:	$y1_t = 130.0u_t$	
Model 2:	$y2_t = 130.0u_t^*$	
Model 2′:	$y2_t = 0.75y2_{t-1} + 85.99u_t$	
Model 3:	$\ln(y3)_t = 0.395\ln(y3)_{t-1} + 0.3995\ln(y3)_{t-2} + 0.00172u_t$	s.e.r. = 0.00172, R^2 =0.99
Model 4:	$y4_t = 1.33x11_t + 9.73u_t$	s.e.r. = 9.73, R^2 = 0.58
Model 5:	$y5_t = -0.046x3_t + 0.11u_t$	s.e.r. = 0.11, R^2 = 0.93
Model 6:	$y6_t = 0.67x11_t – 0.023x3t + 4.92u_t$	s.e.r. = 4.92, R^2 = 0.58
Model 6A:	$y6_t = 0.67x11_t – 0.32x3t + 4.92u_t$	s.e.r. = 4.92, R^2 = 0.64
Model 6B:	$y6_t = 0.67x11_t – 0.65x3t + 4.92u_t$	s.e.r. = 4.92, R^2 = 0.74
Model 7:	$y7_t = 1.33x11_t + 9.73u_t^*$	s.e.r. = 9.73, R^2 = 0.58
Model 7′:	$y7_t = 0.75y7_{t-1} + 1.33x11_t – 0.9975x29_t + 6.73u_t$	
Model 8:	$y8_t = -0.046x3_t + 0.11u_t^*$	s.e.r. = 0.11, R^2 = 0.93
Model 8′:	$y8_t = 0.75y8_{t-1} - 0.046x3_t + 0.00345x21_t + 0.073u_t$	
Model 9:	$y9_t = 0.67x11_t – 0.023x3_t + 4.92u_t^*$	s.e.r. = 4.92, R^2 = 0.58
Model 9′:	$y9_t = 0.75y9_{t-1} – 0.023x3_t + 0.01725x21_t + 0.67x11_t – 0.5025x29_t + 3.25u_t$	

Note: The variables $y\#_t$ are the artificial variables created by each model. The variables $x\#_t$ correspond to the variables with the same number in Table 1. The coefficients for models 3, 4, and 5 come from the regression of personal consumption expenditures (Dep. in Table 1) on independent variables as indicated by the models. The standard error of the regression for models 3, 4, and 5 is scaled to set R^2 equal to that for the analogous regressions run on non-stationary data to mirror Lovell. Model 6 is the average of models 4 and 5. Models 7, 8, and 9 have same coefficients as models 4, 5, and 6 with autoregressive errors. Models 2′, 7′, 8′, and 9′ are exactly equivalent expressions for models 2, 7, 8, 9 in which lags of the variables are used to eliminate the autoregressive parameter in the error process.

4. THE ‘MINING MACHINE’: AN ALGORITHM FOR A GENERAL-TO-SPECIFIC SPECIFICATION SEARCH

The practitioners of the general-to-specific approach usually think of econometrics as an art, the discipline of which comes, not from adhering to recipes, but from testing and running horse-races among alternative specifications. Nevertheless, in order to test the general-to-specific approach in Lovell's framework we are forced to first render it into a mechanical algorithm. The algorithm that we propose is, we believe, a close approximation to a subset of what practitioners of the approach actually do.[12] A number of their concerns, such as appropriate measurement systems and exogeneity status of the variables, are moot because of the way in which we have constructed our nine test models. Also, because we have controlled the construction of the test models in specific ways, considerations of compatibility with economic theory can be left to one side.

[12] See, in addition to the general discussions as indicated in footnote 1 above, Hendry and Richard (1987), White (1990), and Hendry (1995, Ch. 15).

4.1. The search algorithm

A. The data run 1960.3–1995.1. Candidate variables include current and one lag of independent variables and four lags of the dependent variable. A replication is the creation of a set of simulated consumption values using one of the nine models in Table 3 and one draw from the random number generator. Nominal size governs the conventional critical values used in all of the tests employed in the search: it is either 1, 5, or 10%.[13]

B. A general specification is estimated on a replication using the observations from 1960.3 to 1995.1 on the full set of candidate variables, while retaining the observations from 1991.4 to 1995.1 (the 14 observations are 10% of the sample) for out-of-sample testing. The following battery of tests is run on the general specification:

   a. normality of residuals (Jarque and Berra, 1980).
   b. autocorrelation of residuals up to second order (χ^2 test, see Godfrey (1978), Breusch and Pagan (1980)).[14]
   c. autocorrelated conditional heteroscedasticity (ARCH) up to second order (Engle, 1982).
   d. in-sample stability test (first half of the sample against the second half, see Chow (1960)).
   e. out-of-sample stability test of specification estimated against re-estimation using 10% of data points retained for the test Chow (1960).

   If the general specification fails any one of the tests at the nominal size, then this test is not used in subsequent steps of the specification search for the current replication only.[15] If the general specification fails more than one test, the current replication is eliminated and the search begins again with a general specification of a new replication.[16]

C. 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 path begins with the elimination of one of the variables in the subset with the 10 lowest (insignificant) t-statistics as judged by the nominal size. The first search begins by eliminating the variable with the lowest t-statistic and re-estimating the regression. This re-estimated regression becomes the current specification. The search continues until it reaches a terminal specification.

D. Each current specification is subjected to the battery of tests described in step B with the addition of:

   f. An F-test of the hypothesis that the current specification is a valid restriction of the general specification.

[13] A uniform test size is used both for exclusion tests (t-tests) and diagnostic tests. We agree with the suggestion of one referee who believes that it would be worth exploring the effects of independently varying the sizes of the two types of tests.

[14] In using AR(2) and ARCH(2) tests we trade on our knowledge that for every model except model 3, which has a two-period lag, the longest true lag is only one period. As the number of search variables increases with the number of lags, tractability requires some limitation on our models. Given that fact, the limitation of the test statistics to order 2 is probably harmless.

[15] Another and perhaps better option, suggested by a referee, would have been either to use a larger size for the problematic test or to reintroduce the test later in the search. We have, in fact, experimented with both procedures and implemented the second in work-in-progress.

[16] An LSE practitioner would probably prefer in this case to enlarge the general specification, adding variables or lags of existing variables, or to adopt one of the strategies suggested in footnote 15. We drop the specification in this case to facilitate the mechanization of the procedure. In practice, few replications are eliminated this way. For model 7, for instance, only 2 of 1002 replications were eliminated in one run.

E. If the current specification passes all of the tests, the variable with the next lowest t-statistic is eliminated. The resulting current specification is then subjected to the battery of tests. If the current specification fails any one of these tests, the last variable eliminated is restored and the current specification is re-estimated eliminating the variable with the next lowest insignificant t-statistic. The process of variable elimination ends when a current specification passes the battery of tests and either has all variables significant or cannot eliminate any remaining insignificant variable without failing one of the tests.
F. The resultant specification is then estimated over the full sample.
   I. If all variables are significant the current specification is the terminal specification.
   II. If any variables are insignificant, they are removed as a block and the battery of tests is performed.
      a. If the new model passes and all variables are significant the new model is the terminal model and go to G.
      b. If the new model does not pass, restore the block and go to G.
      c. If the new model passes and some variables are insignificant, return to II.
G. After a terminal specification has been reached, it is recorded and the next search path is tried until all 10 have been searched.
H. Once all 10 search paths have ended in a terminal specification, the final specification for the replication is the terminal specification with the lowest standard error of regression.[17]

The general-to-specific search algorithm here is a good approximation to what actual practitioners do, with the exception, perhaps, of the explicit requirement to try several different search paths. We added this feature because preliminary experimentation showed that without it the algorithm frequently got stuck far from any sensible specification. While in this respect our attempt to mechanize LSE econometric methodology may have in fact suggested an improvement to the standard LSE practice, we do not regard this modification as invidious to that practice or as a particularly radical departure. Typically, LSE practitioners regard econometrics as an art informed by both econometric and subject-specific knowledge. We have no way of mechanizing individual econometric craftsmanship. We regard the use of multiple search paths as standing in the place of two normal LSE practices that we are simply unable to model in a simulation study: First, LSE practitioners insist on consistency with economic theory to eliminate some absurd specifications. Since we control the data-generating processes completely, there is no relevant theory to provide an independent check. Second, LSE practitioners typically require that final specifications encompass rival specifications that may or may not have been generated through a general-to-specific search. While the ultimate goal is, of course, to find the truth, the local, practical problem is to adjudicate between specifications that economists seriously entertain as possibly true. We have no set of serious rival specifications to examine. However, if we did, they would no doubt reside at the end of different search paths; so we come close to capturing the relevant practice in considering multiple search paths.[18]

[17] Variance dominance is a necessary condition for encompassing. In work-in-progress we replace this step with an encompassing test of the lowest variance terminal specification against each of the other terminal specifications. If the lowest variance specification fails to encompass any of the other terminal specifications, the non-redundant union of its variables with those of the unencompassed specifications is used as the starting point for a further search. A referee suggested a similar procedure independently.

[18] There may be more than 10 insignificant variables in the general specification. The search algorithm is designed to eliminate any that remain insignificant along the search path unless their retention is needed to pass the test battery. There

5. DOES THE GENERAL-TO-SPECIFIC APPROACH PICK THE TRUE SPECIFICATION?

To assess the general-to-specific approach we conduct a specification search for 1000 replications of each of the nine specifications listed in Table 3. Specifications could be evaluated as either picking out the correct specification or not. We believe, however, that acknowledging degrees of success provides a richer understanding of the efficacy of the search algorithm. We present the results in five categories. Each category compares the *final* specification with the correct or *true* specification that was used to generate the data. The sensibility of the encompassing approach informs the categories. It is a necessary condition that the standard error of regression for an encompassing specification be lower (in population) than every specification that it encompasses. Thus, in population, the true specification must have the lowest standard error of regression. We use this criterion in our search algorithm, but, unfortunately, it need not be satisfied in small samples. We therefore ask: Does the algorithm find the correct model? If not, does it fail because the small sample properties of the data indicate that a rival specification is statistically superior or because the algorithm simply misses? The latter is a serious failure; the former, especially if the true specification, is nested within the final specification, is a near success. We focus on the question of whether or not the true specification is nested within the final specification, because ideally the algorithm would always select the true regressors (i.e. have high power), but is nevertheless subject to type I error (i.e. it sometimes selects spurious additional regressors). The five categories are:

Category 1 (Final = True): *The true specification is chosen*. (The algorithm is an unqualified success.)

Category 2 (True $\subset$ Final, $SER_F < SER_T$):[19] *The true specification is nested in the final specification and the final specification has the lower standard error of regression*. (The algorithm has done its job perfectly, but it is an (adventitious) fact about the data that additional regressors significantly improve the fit of the regression. The final specification appears to encompass the true specification and there is no purely statistical method of reversing that relationship on the available data set.)

Category 3 (True $\subset$ Final, $SER_F > SER_T$): *The true specification is nested in the final specification and the true specification has the lower standard error of regression*. (The algorithm fails badly. Not only does the true specification in fact parsimoniously encompass the final specification, but it could be found if the algorithm had not stopped prematurely on the search path.)

Category 4 (True $\not\subset$ Final, $SER_F < SER_T$): *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*. (The algorithm fails to pick the true specification, but does so for good statistical reasons: given the sample the final specification appears to variance dominate the true specification. It is like category 2 except that, rather than simply including spurious variables, it (also) omits correct variables.)

is nothing sacred about 10 paths; it is an entirely pragmatic choice. We could, as one referee suggested, generate a search path for every insignificant variable or for different blocks of insignificant variables. The simulation data themselves suggest that we would not do substantially better if we considered every possible path: there turn out to be few failures of the algorithm in which the true model dominates the final model. One reason for not trying every path is that to do so would emphasize the mechanical nature of what is in practice not a mechanical procedure.

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

Category 5 (True $\not\subset$ Final, $\mathrm{SER_F} > \mathrm{SER_T}$): *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 than the final specification*. (This is, like category 3, a serious failure of the algorithm—even worse, because the final specification does not even define a class of specifications of which the true specification is one.)

These categories are still too coarse to provide full information about the success of the algorithm. Even category 5 need not always represent a total specification failure. It is possible that a specification may not nest the correct specification but may overlap with it substantially—including some, but not all, of the correct variables, as well as some incorrect variables. We will therefore track for each replication how many times each correct variable was included in the final specifications, as well as the number of additional significant and insignificant variables included.

5.1. A benchmark case: nominal size 5%

Table 4 presents the results of specification searches for 1000 replications of nine specifications for nominal size of 5% (i.e. the critical values based on this size are used in the test battery described in step D of the search algorithm described in Section 4).[20] A 5% size, as the most commonly used by empirical researchers, will serve as our benchmark case throughout this investigation. According to Table 4, the general-to-specific search algorithm chooses exactly the correct specification (category 1) only a small fraction of the time: on average over nine models in 17% of the replications. Its success rate varies with the model: models 1, 3, 4, 5 and 8 give the best results (around 30%), while model 6, 7 and 9 show very low success, and model 2 fails completely to recover the exactly true specification. Still, the general-to-specific algorithm is by no means a total failure. Most of the specifications are classed in category 2, which means that the final specification is overparameterized relative to the true model, but that is the best one could hope to achieve on purely statistical grounds, because the chosen final specification in fact statistically encompasses the true specification. On average 60.7% of searches end in category 2 and nearly 78% in categories 1 and 2 combined. If category 2 is a relative success, the price is overparameterization: an average of just over two extra variables spuriously and significantly retained in the specification. (In addition, in a small number of cases extra insignificant variables are retained.) In one sense, this is bad news for the search algorithm as it suggests that searches will quite commonly include variables that do not correspond to the true data-generating process. But, we can look at it another way. Each falsely included (significant) variable represents a case of type I error. The search is conducted over 40 variables and 1000 replications. The table represents the empirical rate of type I error (size) for the algorithm: on average 6.0%, only a little above the 5% nominal size used in the test battery.

[20]Models 2, 7, 8, 9 involve an AR(1) error term of the form $u_t^* = \rho u_{t-1}^* + u_t$. Each of these models can be expressed as a dynamic form subject to common-factor restrictions. Thus if $y_t = \mathbf{X}_t\boldsymbol{\beta} + u_t^*$, this is equivalent to (a) $y_t = \rho y_{t-1} + \mathbf{X}_t\boldsymbol{\beta} - \mathbf{X}_{t-1}(\rho\boldsymbol{\beta}) + u_t$, so that an estimated regression conforms to (a) if it takes the form (b) $y_t = \pi_1 y_{t-1} + \mathbf{X}_t\boldsymbol{\Pi}_2 - \mathbf{X}_{t-1}\boldsymbol{\Pi}_3 + u_t$, subject to the common-factor restriction $\pi_1\boldsymbol{\Pi}_2 = -\boldsymbol{\Pi}_3$. (NB: bold face symbols represent vectors or matrices.) We present the alternative expressions of the models as models $2'$, $7'$, $8'$ and $9'$. Although many LSE econometricians regard the testing of common-factor restrictions an important element in specification search, we count a search successful if it recovers all the relevant variables (explicit in form (b)), although we do not test the validity of the common-factor restriction itself. See Hoover (1988) and Hendry (1995, Ch. 7, Section 7), for discussions of common-factor restrictions.

Table 4. Specification searches at 5% nominal size.[a]

	True model[b]									
	1[g]	2	3	4	5	6	7	8	9	Means
Percentage of searches for which the true and final specifications are related in categories:[c]										
1. True = Final	29.2	0.0	27.5	29.8	30.2	0.8	4.0	31.6	1.2	17.1
2. True $\subset$ Final, $SER_F < SER_T$	70.6	100.0	65.3	69.9	69.5	7.3	85.7	68.1	9.8	60.7
3. True $\subset$ Final, $SER_F > SER_T$	0.2	0.0	0.1	0.3	0.3	0.0	0.1	0.3	0.1	0.2
4. True $\not\subset$ Final, $SER_F < SER_T$	0.0	0.0	5.9	0.0	0.0	77.1	9.0	0.0	86.5	19.8
5. True $\not\subset$ Final, $SER_F > SER_T$	0.0	0.0	1.2	0.0	0.0	14.8	1.2	0.0	2.4	2.2
True variable number[d]	Null set	37	37/38	11	3	3/11	11/29/37	3/21/37	3/11/21/29/37	
Frequency variables included (percent)	NA	100	98.4/94.5	100	100	8.1/100	100/89.8/100	100/100/100	6.5/100/6.0/ 89.5/100	
Average rate of inclusion per replication of:										
True variables	NA	1.00	1.93	1.00	1.00	1.08	2.90	3.00	3.02	
Insignificant variables	0.28	0.28	0.30	0.27	0.24	0.29	0.40	0.28	0.35	0.3
Falsely significant variables	1.81	4.19	1.87	1.74	1.75	1.59	3.05	1.78	2.97	2.3
Type I error (True Size)[e]	4.5%	10.7%	4.9%	4.5%	4.5%	4.2%	8.2%	3.7%	8.5%	6.0%
Power[f]	N/A	100.0%	96.5%	100.0%	100.0%	54.0%	96.7%	100.0%	60.4%	88.5%

[a]Search algorithm described in text (Section 4). Test batteries use critical values corresponding to two-tailed tests with the nominal size in title. The universe of variables searched over is given in Table 1. All regressions include a constant, which is ignored in evaluation of the successes or failures or searches. Sample runs 1960.3–1995.1 or 139 observations. The table reports the results of 1000 replications. [b]The artificial consumption variable is generated according to the specifications in Table 3. [c]Categories of specification search results are described in the text (Section 5). SER_F indicates the standard error of regression for the final specification and SER_T that for the true specification. [d]Variable numbers correspond to those given in Table 1. [e]Size = falsely significant variables/(total candidates−possible true variables) = relative frequency of rejecting a true null hypothesis. [f]Power = 1 – (possible true variables−true variables chosen)/possible true variables = relative frequency of not accepting a false null hypothesis. [g]For purposes of comparison with the chosen model, the s.e.r. of true is calculated as the standard deviation of y_1.

Again, these averages mask considerable variation across models. At one extreme, almost every search over models 1, 2, 5, and 8 ends in category 1 or 2. At the other extreme only about 10% of searches over model 6 and 9 end in categories 1 or 2. For models 3 and 7, a substantial proportion of searches end in categories 1 and 2, but a smaller, though not insignificant number, end in categories 4 and 5, which are more serious failures of the algorithm. So, how do these models fail?

5.2. Weak signals, strong noise

Searches for both models 6 and 9 most frequently end in category 4: the true specification is not nested within the final specification, but the final specification (statistically) variance dominates the true specification. This suggests, not a failure of the algorithm, but unavoidable properties of the data. Table 4 indicates that models 6 and 9 correctly choose most of the true variables most of the time, but that they appear to have special difficulty in capturing government purchases of goods and services (Variable 3) or its first lagged value (Variable 21). We conjecture that the difficulty in this case is that these variables have relatively low variability compared with the dependent variables and the other true independent variables in models 6 and 9. They therefore represent a common and unavoidable econometric problem of variables with a low signal-to-noise ratio.[21] It is always problematic how to discriminate between cases in which such variables are economically unimportant and cases in which they are merely hard to measure.

Consider model 6 in more detail. The signal-to-noise ratio for variable j in the true model can be defined as $S_j = |\beta_j \sigma_j|/\sigma_\varepsilon$, where β_j is the true coefficient for independent variable j, σ_j is the standard deviation of independent variable j, and σ_ε is the standard deviation of the random error term for the model. In model 6, the signal-to-noise ratio for Variable 3 is $S_3 = 0.04$, while for Variable 11 (the M1 monetary aggregate) $S_{11} = 1.16$. By adjusting β_3, S_3 can be increased. We formulate two additional models (6A and 6B) in which β_3 is raised (in absolute value) from -0.02 to -0.32 and then to -0.67, yielding signal-to-noise ratios of 0.58 (half of that for Variable 11) and 1.16 (the same as that for Variable 11). Table 5 presents the results of 1000 replications of the search at a nominal size of 5% for models 6, 6A, and 6B. With even half the signal-to-noise ratio of Variable 11, the final specification for model 6A ends up 86.2% of the searches in categories 1 or 2, and Variable 3 is correctly selected in 86.4% of those searches. With an equal signal-to-noise ratio, the final specification for model 6B ends up with nearly 100% of the searches in categories 1 and 2, and Variable 3 is selected correctly in almost every case.

5.3. Size and power

How do the properties of the general-to-specific search algorithm change as the nominal size used in the test battery changes? Tables 6 and 7 present analogous results to those in Table 4 (nominal size 5%) for nominal sizes of 10% and 1%. Some general patterns are clear in comparing the three

[21]The reader will notice that in models 5 and 8, these variables appear to present no special difficulties. There is, however, no paradox. The relevant factors are not only the absolute variability of the dependent variable, but also the size of the coefficient that multiplies it; and these must be judged relative to the other independent variables in the regression, as well as to the dependent variable (and therefore, finally, to the error term). The fact that these variables are easily picked up in cases in which there are no competing variables merely underlines the fact that it is the *relative* magnitudes that matter.

tables. As the nominal size falls, the number of final specifications in category 1 rises sharply from an average of under 5% at a nominal size of 10% to an average of nearly 50 at a nominal size of 1%. At the same time, the relationship between nominal size and category 2 is direct not inverse, and the total in categories 1 and 2 together is lower (average almost 75%) for a nominal size of 1% than for a nominal size of 5% (nearly 78%) or 10% (just over 80%). Similarly, a smaller nominal size sharply reduces the average number of both falsely significant variables and retained insignificant variables. All these features are indications of the tradeoff between size and power. The average true size corresponding to a 10% nominal size is 11.6%—almost identical—and is associated with an average power of 89.3%. The true size corresponding to a nominal size of 5% is also close, 6.0%, but the reduction in size implies a slight loss of power (down to 88.5%). The smaller size implies fewer cases of incorrectly chosen variables, but more cases of omitted correct variables. The true size corresponding to a nominal size of 1% is almost double at 1.8%, and there is a further loss of power to 87.0%. The tradeoff between size and power seems to be pretty flat, although as nominal size becomes small the size distortion becomes relatively large. This may argue for a smaller conventional size in practical specification searches than the 5% nominal size commonly used (Hendry, 1995, p. 491).

Table 5. Specification searches at 5% nominal size.[a]

	True model[b]		
	6	6A	6B
Percentage of searches for which the true and final specifications are related in categories: [c]			
1. True = Final	0.8	27.4	33.1
2. True $\subset$ Final, $SER_F < SER_T$	7.3	58.8	66.1
3. True $\subset$ Final, $SER_F > SER_T$	0.0	0.1	0.1
4. True $\not\subset$ Final, $SER_F < SER_T$	77.1	11.1	0.5
5. True $\not\subset$ Final, $SER_F > SER_T$	14.8	2.6	0.2
True variable number[d]	3/11	3/11	3/11
Variable included (percent)	8.1/100	86.4/99.9	99.6/99.7
Average rate of inclusion per replication of:			
True variables	1.08	1.86	1.99
Insignificant variables	0.29	0.20	0.24
Falsely significant variables	1.59	1.89	1.65
Type I error (true size)[e]	4.2%	4.8%	4.3%
Power[f]	54.0%	93.0%	99.5%

[a]Search algorithm described in text (Section 4). Test batteries use critical values corresponding to two-tailed tests with the nominal size in title. The universe of variables searched over is given in Table 1. All regressions include a constant, which is ignored in evaluation of the successes or failures or searches. Sample runs 1960.3–1995.1 or 139 observations. The table reports the results of 1000 replications. [b]The artificial consumption variable is generated according to the specifications in Table 3.[c]Categories of specification search results are described in the text (Section 5). SER_F indicates the standard error of regression for the final specification and SER_T that for the true specification. [d]Variable numbers correspond to those given in Table 1. [e]Size = falsely significant variables/(total candidates – possible true variables) = relative frequency of rejecting a true null hypothesis. [f]Power = 1 – (possible true variables – true variables chosen)/possible true variables = relative frequency of not accepting a false null hypothesis.

Table 6. Specification searches at 10% nominal size.[a]

	True model[b]									
	1[g]	2	3	4	5	6	7	8	9	Means
Percentage of searches for which the true and final specifications are related in categories:[c]										
1. True = Final	7.0	0.0	7.9	8.4	7.7	0.1	0.2	7.6	0.4	4.37
2. True $\subset$ Final, $SER_F < SER_T$	92.9	100.0	86.9	91.4	92.1	14.9	90.3	91.4	19.9	75.64
3. True $\subset$ Final, $SER_F > SER_T$	0.1	0.0	0.4	0.1	0.2	0.0	0.2	1.0	0.0	0.22
4. True $\not\subset$ Final, $SER_F < SER_T$	0.0	0.0	4.3	0.1	0.0	81.3	9.0	0.0	79.4	19.34
5. True $\not\subset$ Final, $SER_F > SER_T$	0.0	0.0	0.5	0.0	0.0	3.7	0.3	0.0	0.3	0.53
True variable number[d]	Null set	37	37/38	11	3	3/11	11/29/37	3/21/37	3/11/21/29/37	
Frequency variables included (percent)		100.0	98.3/96.9	99.9	100.0	15.0/99.9	100.0/90.7/100.0	100.0/100.0/100.0	11.9/100.0/10.8/89.7/100.0	
Average rate of inclusion per replication of:										
True variables		1.00	1.95	0.99	1.00	1.15	2.91	3.00	3.12	
Insignificant variables	0.64	0.67	0.68	0.60	0.65	0.51	0.82	0.70	0.70	0.66
Falsely significant variables	3.99	6.30	3.84	3.83	3.85	3.63	5.26	3.90	4.95	4.39
Type I error (true size)[e]	10.0%	16.2%	10.1%	9.8%	9.9%	9.5%	14.2%	10.6%	14.1%	11.6%
Power[f]	N/A	100.0%	97.6%	99.9%	100.0%	57.5%	96.9%	100.0%	62.5%	89.3%

[a]Search algorithm described in text (Section 4). Test batteries use critical values corresponding to two-tailed tests with the nominal size in title. The universe of variables searched over is given in Table 1. All regressions include a constant, which is ignored in evaluation of the successes or failures or searches. Sample runs 1960.3–1995.1 or 139 observations. The table reports the results of 1000 replications. [b]The artificial consumption variable is generated according to the specifications in Table 3.[c]Categories of specification search results are described in the text (Section 5). SER_F indicates the standard error of regression for the final specification and SER_T that for the true specification.[d]Variable numbers correspond to those given in Table 1. [e]Size = falsely significant variables/(total candidates – possible true variables) =relative frequency of rejecting a true null hypothesis. [f]Power = 1 – (possible true variables – true variables chosen)/possible true variables = relative frequency of not accepting a false null hypothesis. [g]For purposes of comparison with the chosen model, the s.e.r. of true is calculated as the standard deviation of y_1.

Table 7. Specification search at 1% nominal size.[a]

	True model[b]									
	1[g]	2	3	4	5	6	7	8	9	Means
Percentage of searches for which the true and final specifications are related in categories:[c]										
1. True = Final	79.9	0.8	70.2	80.2	79.7	0.7	24.6	78.0	0.8	46.1
2. True $\subset$ Final, $SER_F < SER_T$	20.1	99.2	19.0	19.6	20.2	0.1	57.4	21.7	1.3	28.7
3. True $\subset$ Final, $SER_F > SER_T$	0.0	0.0	0.2	0.1	0.1	0.0	0.0	0.2	0.6	0.1
4. True $\not\subset$ Final, $SER_F < SER_T$	0.0	0.0	3.7	0.1	0.0	56.3	13.0	0.1	77.0	16.7
5. True $\not\subset$ Final, $SER_F > SER_T$	0.0	0.0	6.9	0.0	0.0	42.9	5.0	0.0	20.3	8.3
True variable number[d]	Null set	37	37/38	11	3	3/11	11/29/37	3/21/37	3/11/21/29/37	
Frequency variables included (percent)		100.0	95.7/93.6	99.9	100.0	0.8/99.8	100.0/82.0/ 100.0	100.0/99.9/ 99.9	1.5/100.0/ 1.4/83.5/99.9	
Average rate of inclusion per replication of:										
True variables	N/A	1.00	1.89	0.99	1.00	1.01	2.82	3.00	2.86	
Insignificant variables	0.01	0.07	0.04	0.05	0.04	0.02	0.11	0.05	0.06	0.05
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)[e]	0.7%	5.7%	0.9%	0.8%	0.7%	0.6%	3.0%	0.9%	3.2%	1.8%
Power[f]	N/A	100.0%	94.7%	99.9%	100.0%	50.3%	94.0%	99.9%	57.3%	87.0%

[a]Search algorithm described in text (Section 4). Test batteries use critical values corresponding to two-tailed tests with the nominal size in title. The universe of variables searched over is given in Table 1. All regressions include a constant, which is ignored in evaluation of the successes or failures or searches. Sample runs 1960.3–1995.1 or 139 observations. The table reports the results of 1000 replications. [b]The artificial consumption variable is generated according to the specifications in Table 3. [c]Categories of specification search results are described in the text (Section 5). SER_F indicates the standard error of regression for the final specification and SER_T that for the true specification. [d]Variable numbers correspond to those given in Table 1. [e]Size = falsely significant variables/(total candidates – possible true variables) =relative frequency of rejecting a true null hypothesis. [f]Power = 1 – (possible true variables – true variables chosen)/possible true variables = relative frequency of not accepting a false null hypothesis. [g]For purposes of comparison with the chosen model, the s.e.r. of true is calculated as the standard deviation of y_1.

6. WHAT DO TEST STATISTICS MEAN AFTER EXTENSIVE SEARCH?

The most common doubt expressed about the final specifications reported from general-to-specific specification searches is over the interpretation of test statistics. How are we to interpret the t-statistics of a regression that involves massive (and not easily quantified) amounts of pre-test selection and (it is pejoratively but wrongly argued) arbitrarily directed search? Should we not, following Lovell for example, discount the test statistics in proportion to the degree of search? It would be desirable to be assured that an algorithm converged on the true data-generating process. In that case, the sampling properties of the final specification would be the sample properties of the true specification. The results of the previous section, however, indicate a number of pitfalls that might vitiate the success of the general-to-specific algorithm. It is only relatively infrequently that it converges on the exactly correct specification. Commonly, a relatively large number of extra significant regressors are included in the final specification, and extra insignificant regressors are often apparently needed to obtain desirable properties for the estimated residuals. In the face of these common departures from a precise match between the chosen final specifications and the true specification, the question posed in this section is, to what degree does the final specification reflect the sampling properties of the true specification?

To investigate this question we conduct specification searches on 1000 replications of model 9. Model 9 was chosen because it is the most difficult of Lovell's nine models for the search algorithm to uncover. It is both a dynamic model and one that suffers from low signal-to-noise ratios for some of its variables. Table 8 presents the results of this exercise for the universe of variables in Table 1 for searches with a nominal size of 5%.

Although every variable in the universe of search is chosen in some replications and therefore have non-zero mean values, incorrect inclusion is relatively rare. This is highlighted by the fact that the median values of the correctly excluded coefficients are almost always zero. A more detailed examination of the individual variables than is shown in Table 8 indicates that only Variable 38, the second lag of the dependent variable (artificial consumption expenditures), has a non-zero median. It is chosen (incorrectly) in nearly 88% of the replications, while its brother, the (correct) first lag (Variable 37), is chosen in nearly 100% of the replications, so that in most cases both variables are chosen. We will return to this phenomenon presently.

Concentrating now on the properly included variables, we measure the accuracy of the estimates as the absolute values of the mean and median coefficient biases as a percentage of the true value. Variable 11 appears to be fairly accurately measured with mean bias of 2.4% and median bias of 3.1%. The biases of Variables 29 and 37 are substantially higher but still moderate. In contrast, the two variables with low signal-to-noise ratios (Variables 3 and 21) have very large mean biases of 107% and 75% and median biases of 100%.

To evaluate the interpretation of t-statistics, we kept track of the estimated t-statistics for each final specification. We measured the type I error for the properly excluded variables as the number of times that the t-statistic was outside the 95% confidence interval (i.e. the number of times a variable was improperly included with $|t\text{-statistic}| > 1.96$) and the type II error for the properly included variables as the number of times the t-statistic was inside the 95% confidence interval. From these data we can compute the empirical size and power of the t-test against the null hypothesis that the coefficient on a variable is zero (exclusion of a variable from the search is treated as being equivalent to a coefficient value of zero).

The empirical sizes of the properly excluded variables average about 8.5%. Variable 38 is the second lagged value of the dependent variable. This variable, as we noted previously, is the only variable that is incorrectly chosen more often than not. It is highly correlated with the first lag of the

Table 8. Monte Carlo statistics for specification search on model 9 (1000 replications).

	Variables						
	Correctly included					Correctly excluded	
	3	11	21	29	37	All	All except 38, 39 and 40
True value[a]	−0.023	0.670	0.017	−0.500	0.750	0.000	0.000
Estimated coefficients							
Mean	0.002	0.686	0.004	−0.294	0.574	0.004	0.010
Median	0.000	0.691	0.000	−0.322	0.578	−0.007	0.000
Max	0.329	0.960	0.166	0.000	0.859	1.476	1.599
Min	−0.308	0.307	−0.137	−0.611	0.000	−1.246	−1.334
Standard deviation	0.044	0.091	0.027	0.140	0.106	0.237	0.252
Simulated standard deviation[b]	0.05	0.06	0.05	0.07	0.06		
Mean bias[c] (percent)	106.7	2.4	75.0	41.3	23.5		
Median bias[d] (percent)	100.0	3.1	100.0	35.6	22.9		
Empirical size[e] (percent)						8.5	5.4
True power[f] (percent)	10.0	100.0	8.0	100.0	100.0		
Empirical power[g] (percent)	9.4	100.0	9.1	86.0	99.7		
Chosen but insignificant (percent)						3.7	3.8

[a]Coefficients from model 9′, Table 3. [b]Actual standard deviation of coefficients from 1000 replications of model 9 (i.e. without search). [c]|(mean estimated values – true value)|/true value expressed as percentage. [d]|(median estimated values – true value)|/true value expressed as percentage. [e]Proportion of t-statistics outside ±1.96 (i.e., the nominal 5 percent critical value). [f]Proportion of t-statistics inside ±1.96 (i.e., the nominal 5 percent critical value) for 1000 replications of model 9′ (i.e. without search). [g]Proportion of t-statistics inside ±1.96 (i.e., the nominal 5 percent critical value).

dependent variable (correlation coefficient 0.75).[22] This multicollinearity is the likely source of the large empirical size. While we should regard this example as a warning of one of the pitfalls of dynamic specification search, it may say more about the inadequacy of our algorithm in mimicking the recommended practice of the LSE approach. The LSE methodology stresses the importance of *orthogonal* regressors and the need to find reparameterizations to ensure orthogonality. If we do not count the three properly excluded lags of the dependent variable (Variables 38, 39, and 40), then the average empirical size for the remaining properly excluded variables is 5.4%, very close to the nominal size of 5% used in the search algorithm.

Since we know the true specification of model 9, it is possible to compute the power against the null that the coefficient on any properly included variable is zero for any single replication. In order to account for the fact that the dependent variable (and its lagged value) varies with each replication, we compute the power from 1000 replications and estimates of the true model. This is indicated in Table 10 as the 'true power'. We compare the estimated empirical power of the search algorithm against this true power. While the empirical power varies tremendously with the variable (100% for Variable 11 but just over 9% for Variable 21), there is a close conformity between the empirical power and the true power. The largest discrepancy occurs with Variable 29

[22]The correlation is measured using actual personal consumption expenditure rather than the simulated dependent variable, which varies from replication to replication. The correlation should be close in any case.

(the first lag of Variable 11, the M1 monetary aggregate), which has an empirical power of 86% against a true power of 100%. Once again this may be the result of the high correlation between the current and lagged values of the variable (correlation coefficient = 0.682).

In summary, the size and power of final specifications from the general-to-specific search algorithm provide very good approximations to the size and power of the true specifications. We have also conducted, but do not report here, two further sets of 1000 replications for nominal sizes of 10% and 1%. The results are similar in character to those in Table 8.

7. THE PROBLEM OF OVERFITTING: AN EXTENSION TO THE LSE METHODOLOGY

Our investigations confirm the worry of some critics who believe that the general-to-specific search results in overparameterized models. Final specifications, more often than not, retain incorrectly significant variables and, less frequently, insignificant variables that appear to be needed to induce sensible properties in the error terms. Given that we have shown that the empirical size and power of t-tests are not very distorted by the search procedure, this is perhaps of less concern than it first appears. Furthermore, the problem appears to be substantially mitigated through the use of smaller nominal sizes in the search procedure. We have shown that the cost of using smaller nominal sizes in terms of power is relatively small. Thus, as well as evaluating the LSE approach, we make a constructive suggestion that practitioners should prefer smaller nominal test sizes.

Type I error in the search process occurs because the data possess adventitious properties in small samples. By their very nature these properties should not remain stable across subsamples. This suggests a possible method of reducing the number of incorrectly retained significant variables (i.e. reducing the empirical size of the algorithm), which, to the best of our knowledge, is not generally practiced by LSE econometricians, but which is consistent with the general philosophy of the LSE methodology. We consider splitting the sample into two (possibly overlapping) subsamples—one running from the beginning of the sample to a point some fraction of the way to the end, the second running from the end of the sample some fraction of the way backwards to the beginning. If, for example, the fraction is one half, the subsamples are the first half and the second half of the full sample, and they do not overlap. If the fraction is 60%, the subsamples are the first 60 and the last 60% of the full sample; the two subsamples overlap in the middle 20% of the full sample. We run a modified version of the search algorithm on each subsample. The final model is then the intersection of the two subsample models; that is, only variables that are chosen in both subsamples appear in the final model, on the grounds that the others are there by accidents of the data.[23]

The algorithm of Section 3 above is modified by omitting step B.d, the in-sample Chow test for coefficient stability and reducing the number of data points retained for out-of-sample stability testing in step B.e (maintaining the 10% ratio). Both modifications are pragmatic responses to the loss of degrees of freedom from the use of shorter subsamples.

[23]While we believe that no LSE econometrician has proposed this precise procedure, it is related to their common use of recursive regressions and diagnostics based on them (see, for example, Doornik and Hendry (1997, pp. 95–97), who considered recursive tests in the context of specifying parsimonious VARs in PC-Fiml). Ericsson (1998, p. 87) comes close to our proposal with the suggestion that a recursive t-statistic that peaks in midsample rather than rising across the entire sample is symptomatic of adventitious correlation. Test based on recursive regressions are, unfortunately, difficult to render into a mechanical algorithm.

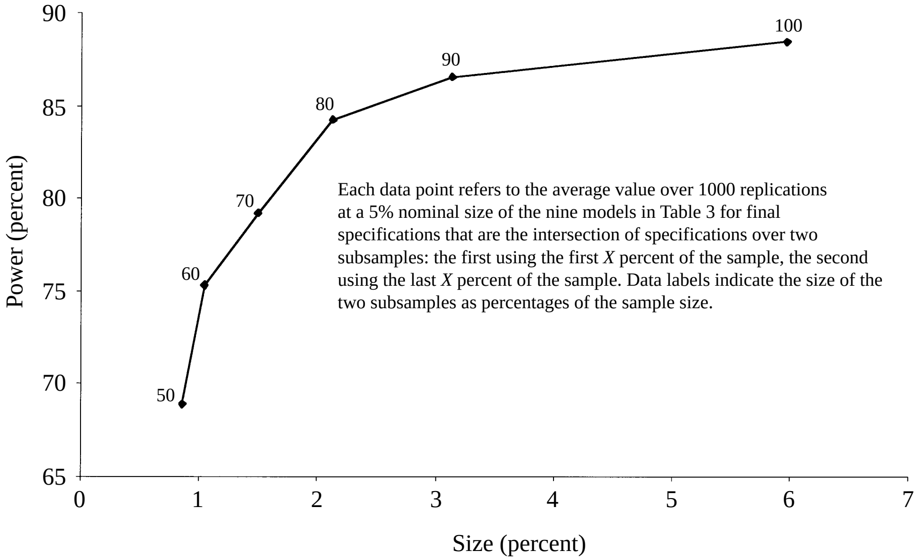

Figure 1. Average size–power tradeoff for split sample searches.

For 1000 replications of the nine models with subsamples of one half the data set, the average number of falsely included significant variables is 0.30 compared with 2.3 for the full data set in Table 4. This is a fall in the empirical size to 0.9% from 6.0%. The improvement in size, however, comes at the cost of great loss of power: 68.9% compared with 88.5% for the full sample.

Figure 1 plots the tradeoff between size and power for subsamples consisting of increasingly large fractions of the whole sample based on 1000 replications of the nine models. The tradeoff is non-linear: the highest power occurs naturally with the full undivided sample; the loss of power is relatively small up to the point at which the subsamples are 80% of the full sample and then falls rapidly to the point where the subsamples are half the full sample. The tradeoff locus can be regarded as a possibility frontier, and an investigator's loss function would rank the various possibilities (higher indifference curves would lie to the northwest). Obviously, any of the points along the locus is a conceivable optimum. Still, for a large class of loss functions the kink at the 80% subsample would prove to be the optimum. At that point the average size is 2.1% (about a third of the size reported in Table 4), and the average power is 84.3% (a loss of only 4.2 percentage points or about 4.7% compared with the power reported in Table 4). With a well-chosen subsample split, the modified algorithm produces a large improvement in size (reduction in overparameterization) for a small loss of power.

8. DATA MINING IN RETROSPECT. . . AND PROSPECT

The results of our investigation of the general-to-specific search algorithm should be reasonably heartening to practitioners of the LSE approach. Unlike Lovell (1983), we find that the general-to-specific approach recovers the correct specification or a closely related specification most of the time. Furthermore, the empirical size and power of specifications produced from general-to-specific searches, with one caveat, conform well to the theoretical size and power one would expect if one knew—and knew that one knew—the true specification *a priori*. Test statistics based on such searched specifications therefore bear the conventional interpretation one would ascribe to one-shot tests. Of course, estimated standard errors are measures of sampling characteristics, not of epistemic virtue. This remains true with a searched specification. A t-statistic may be insignificant either because a variable is economically unimportant or because it has a low signal-to-noise ratio or small sample. The searched specification may, nevertheless possess epistemic virtues not open to the one-shot test: since the correct specification necessarily encompasses all incorrect specifications, the fact that the searched specification is naturally nested within a very general specification, which nests a wide class of alternative specifications in its turn, strengthens the searched specification as a contender for the place of model-most-congruent-to-the-truth. The evidence of strength is not found in the t-statistics, but in the fact of the Darwinian survival of the searched specification against alternatives and in its natural relationship to the general specification.

The one caveat is that our evidence shows that size certainly and, to a lesser extent, power are distorted for lags of (especially, the dependent) variables of the true specification. This appears to be concerned with failures of orthogonality. At a minimum, it reminds the practitioner why the LSE approach stresses the importance of orthogonality and special care with respect to dynamic specification.

While generally supportive of the LSE approach, this study was able to confirm the risk often asserted by critics that practical general-to-specific searches could turn into arbitrary wanderings in the maze of specification possibilities that might terminate arbitrarily far from the correct specification. While the LSE approach in fact incorporates a number of elements (ignored in our mechanical rendering of the search procedure) that protect against false termini, we found that the simple expedient of trying a number of initial starting points in the search gave very good results. We recommend this to practitioners.

Finally, we would like to pursue two further extensions of the current study. First, we have restricted the models to stationary data. In the past decade, it has become increasingly important in macroeconometrics to deal with non-stationary data. Practitioners of the LSE approach were early contributors to this development, stressing the importance of error-correction modeling long before cointegration had been named or its intimate relationship to error-correction models understood. It is, therefore, natural that we should attempt to evaluate the success of the general-to-specific approach in non-stationary contexts.

Finally, an important alternative view of specification is provided by Leamer (1983, 1985). Leamer regards specification search as inevitable and makes a particular proposal, 'extreme-bounds analysis,' to guide practitioners on the epistemic virtues of estimated regressions. It would be useful to conduct a detailed comparison of the two approaches.[24]

9. ACKNOWLEDGEMENTS

We thank Neil Ericsson, Jon Faust, Clinton Greene, James Hartley, David Hendry, Edward Leamer, Michael Lovell, Thomas Mayer, Steven Sheffrin, Neil Shephard, the participants in workshops and seminars at the University of California, Davis, the University of Amsterdam,

[24]There are already several articles critical of Leamer's approach from an LSE perspective; see, for example, McAleer *et al.* (1983) (and Leamer's (1985) reply), Mizon and Hendry (1990), and Pagan (1987). In work-in-progress, we investigate the relative performance of two modifications of Leamer's approach that have been applied to cross-country studies of the determinants of differences in GDP growth rates (Levine and Renelt, 1992 and Sala-i-Martin, 1997).

Virginia Commonwealth University, and the Board of Governors of the Federal Reserve System, as well as two anonymous referees for helpful comments on earlier drafts.

REFERENCES

Baba, Y., D.F. Hendry and R. M. Starr (1992). The demand for M1 in the U.S.A. *Review of Economic Studies 59*, 25–61.

Banerjee, A. (1995). Dynamic specification and testing for unit roots and cointegration. In K. D. Hoover (ed.), *Macroeconometrics: Developments, Tensions and Prospects*, pp. 473–500. Boston: Kluwer.

Breusch, T. S. and A.R . Pagan. (1980).The Lagrange multiplier test and its application to model specification in econometrics, *Review of Economic Studies 47*, 239–53.

Campbell, J. Y. and P. Perron (1991). Pitfalls and opportunities: What macroeconomists should know about unit roots. In O. J. Blanchard and S. Fischer (eds), *NBER Macroeconomics Annual 1991*, pp. 141–201. Cambridge, MA: MIT Press.

Canova, F. (1995). The economics of VAR models. In K. D. Hoover (ed.), *Macroeconometrics: Developments, Tensions and Prospects*, pp. 57–98. Boston, MA: Kluwer.

Chatfield, C. (1995). Model uncertainty, data mining and statistical inference. *Journal of the Royal Statistics Society, Series A 158*, 419–66.

Chow, G. (1960).Tests of equality between sets of coefficients in two linear regressions, *Econometrica* 28, 591–605.

Cox, D. R. (1982). Statistical significance tests. *British Journal of Clinical Pharmacology 14*, 325–31.

Dolado, J., T. J. Jenkinson and S. S. Rivero (1990). Cointegration and unit roots. *Journal of Economic Surveys 4*, 249–73.

Doornik, J.A. and D. F. Hendry. (1997). *Modelling Dynamic Systems Using PC-Fiml 9.0 for Windows*. London: International Thompson Business Press.

Engle, R. F. (1982). Autoregressive conditional heteroscedasticity, with estimates of the variance of United Kingdom inflations, *Econometrica 50*, 987–1007.

Ericsson, N. R. (1998). Course lecture notes for empirical modeling of macroeconomic time-series. Parts 2 and 3. Washington, D.C.: IMF Institute, International Monetary Fund.

Ericsson, N. R., J. Campos and H. A. Tran (1990). PC-GIVE and David Hendry's econometric methodology. *Revista de Econometrica 10*, 7–117.

Faust, J. and C. H. Whiteman (1995). Commentary [on Grayham E. Mizon's Progressive modeling of macroeconomic times series: The LSE methodology]. In K. D. Hoover (ed.), *Macroeconometrics: Developments, Tensions and Prospects*, pp. 171–180. Boston, MA: Kluwer.

Faust, J. and C. H. Whiteman (1997). General-to-specific procedures for fitting a data-admissible, theory-inspired, congruent, parsimonious, encompassing, weakly-exogenous, identified, structural model to the DGP: A translation and critique, *Carnegie-Rochester Conference Series on Economic Policy 47*, 121–62.

Gilbert, C. L. (1986). Professor Hendry's econometric methodology. *Oxford Bulletin of Economics and Statistics 48*, 283–307.

Godfrey, L.G. (1978). Testing for higher order serial correlation in regression equations when the regressors include lagged dependent variables, *Econometrica 46*, 1303–13.

Hansen, B. E. (1996). Methodology: Alchemy or science? *Economic Journal 106*, 1398–1431.

Hendry, D. F. (1987). Econometric methodology: A personal viewpoint. In T. Bewley (ed.), *Advances in Econometrics*, vol. 2. Cambridge: Cambridge University Press, pp. 29–48.

Hendry, D. F. (1988). Encompassing. *National Institute Economic Review*, August, 88–92.

Hendry, D. F. (1995). *Dynamic Econometrics*. Oxford: Oxford University Press.

Hendry, D. F. (1997). On congruent econometric relations: A comment. *Carnegie–Rochester Conference Series on Public Policy 47*, 163–90.

Hendry, D. F. and J.-F. Richard. (1987). Recent developments in the theory of encompassing. In B. Cornet and H. Tulkens (eds), *Contributions to Operations Research and Economics: The Twentieth Anniversary of Core*, pp. 393–440. Cambridge, MA: MIT Press.

Hess, G. D., C. S. Jones and R. D. Porter (1998). The predictive failure of the Baba, Hendry and Starr model of M1. *Journal of Economics and Business 50*, 477–507.

Hoover, K. D. (1988). On the pitfalls of untested common-factor restrictions: The case of the inverted Fisher hypothesis. *Oxford Bulletin of Economics and Statistics 50*, 135–39.

Hoover, K. D. (1995). In defense of data mining: Some preliminary thoughts. In K. D. Hoover and S. M. Sheffrin (eds), *Monetarism and the Methodology of Economics*: Essays in Honour of Thomas Mayer. Aldershot: Edward Elgar, pp. 242–57.

Ingram, B. F. (1995). Recent advances in solving and estimating dynamic macroeconomic models. In K. D. Hoover (ed.), *Macroeconometrics: Developments, Tensions and Prospects*, pp. 15–46. Boston, MA: Kluwer.

Jarque, C. M. and A. K Berra (1980). Efficient tests for normality, homoscedasticity and serial independence of regression residuals. *Economic Letters 6*, 255–59.

Kydland, F. E. and E. C. Prescott (1995). The econometrics of the general equilibrium approach to business cycles. In K. D. Hoover (ed.), *Macroeconometrics: Developments, Tensions and Prospects*, pp. 181–98, Boston, MA: Kluwer.

Leamer, E. (1978). *Specification Searches: Ad Hoc Inference with Nonexperimental Data*. Boston: John Wiley.

Leamer, E. (1983). Let's take the con out of econometrics. *American Economic Review 73*, 31–43.

Leamer, E. (1985). Sensitivity analysis would help. *American Economic Review 75*, 308–13.

Levine, R. and D. Renelt (1992). A sensitivity analysis of cross-country growth regressions. *American Economic Review 82*, 942–63.

Lovell, M. C. (1983). Data mining. *Reveiw of Economic Statistics 65*, 1–12.

Mayer, T. (1980). Economics as a hard science: Realistic goal or wishful thinking? *Economic Inquiry 18*, 165–78.

Mayer, T. (1993). *Truth versus Precision in Economics*. Aldershot: Edward Elgar.

McAleer, M., A. R. Pagan and P. A. Volker (1983). What will take the con out of econometrics? *American Economic Review 75*, 293–307.

Mizon, G. E. (1984). The encompassing approach in econometrics. In D.F. Hendry and K.F. Wallis (eds), *Econometrics and Quantitative Economics*, pp. 135–72. Oxford: Blackwell.

Mizon, G. E. (1995). Progressive modelling of macroeconomic time series: The LSE methodology. In K. D. Hoover (ed.), *Macroeconometrics: Developments, Tensions and Prospects*, pp. 107–70. Boston: Kluwer.

Mizon, G. E. and D. F. Hendry (1990). Procrustean econometrics: Or stretching and squeezing data. In Granger, C. W. J. (ed.) (1990). *Modelling Economic Series: Readings in Econometric Methodology*, pp. 121–36. Oxford: Clarendon Press.

Mizon, G. E. and J.-F. Richard (1986). The encompassing principle and its application to testing non-nested hypotheses. *Econometrica 54*, 657–78.

Nester, M. R. (1996). An applied statistician's creed. *Applied Statistics 45*, 401–10.

Pagan, A. (1987). Three econometric methodologies: A critical appraisal. *Journal of Economic Surveys 1*, 3–24.

Phillips, P. C. B. (1988). Reflections on econometric methodology. *Economic Record 64*, 334–59.

Phillips, P. C. B. and P. Perron (1988) Testing for a unit root in time series regression. *Biometrika 73*, 355–46.

Sala-i-Martin, X. (1997). I just ran two million regressions. *American Economic Review 87*, 178–83.

Stock, J. H. and M. W. Watson (1988). Variable trends in economic times series. *Journal of Economic Perspectives 2*, 147–74.

White, H. (1990). A consistent model selection procedure based on m-testing. In Granger, C. W. J. (ed.) (1990). *Modelling Economic Series: Readings in Econometric Methodology*, pp. 369–83. Oxford: Clarendon Press.