

Improving on ‘Data mining reconsidered’ by K.D. Hoover and S.J. Perez

DAVID F. HENDRY AND HANS-MARTIN KROLZIG

Nuffield College, Oxford, OX1 1NF, UK

E-mail: david.hendry@nuffield.ox.ac.uk;

Homepage: www.economics.ox.ac.uk/hendry

Economics Department, Oxford University, OX1 3UL, UK

E-mail: hans-martin.krolzig@nuffield.ox.ac.uk;

Homepage: www.economics.ox.ac.uk/hendrykrolzig

Received: November 1999

Summary Kevin Hoover and Stephen Perez take important steps towards resolving some key issues in econometric methodology. They simulate general-to-specific selection for linear, dynamic regression models, and find that their algorithm performs well in re-mining the ‘Lovell database’. We discuss developments that improve on their results, automated in *PcGets*. Monte Carlo experiments and re-analyses of empirical studies show that pre-selection F-tests, encompassing tests, and sub-sample reliability checks all help eliminate ‘spuriously-significant’ regressors, without impugning recovery of the correct specification.

Keywords: *Econometric methodology, Model selection, Encompassing, Data mining, Monte Carlo experiments, Money demand.*

1. INTRODUCTION

In evaluating general-to-specific model selection (denoted *Gets*: see e.g., Hendry (1995, chs. 8 and 14)), Hoover and Perez (1999) introduce some important innovations relevant to all empirical modelling approaches, thereby facilitating a range of further enhancements. They objectively examine a central methodological issue, and are unusually clear on most issues of how to interpret evidence after data-based searches, including whether statistical or epistemic aspects are changed by searching. Thus, it is a pleasure to comment on their paper.

While Hoover and Perez find that computer-automated *Gets* performs well in a range of Monte Carlo experiments, that became true only after they implemented multiple-path searches. Moreover, there remain experiments on which their approach does not do well, given the criteria they used. Thus, we suggest alternative evaluation standards for such experiments, and have written an algorithm that significantly improves on their work – but which would not have been envisaged that way prior to their paper. We will discuss the innovations they have introduced, and how those can be further improved, then briefly explain our own algorithm, before demonstrating its performance both by re-running their Monte Carlo simulations, and re-analysing a well-known empirical model.

Our paper first describes their experimental design, and graphs the data variables used. We then draw a distinction between the unavoidable costs of inference – those which must arise when

the ‘truth’ is unknown, so tests have to be conducted to determine the relevant variables – and the (possibly avoidable) costs of search which arise when trying to ascertain the causal variables from a candidate set that is inadvertently too large. This distinction leads us to propose a different evaluation standard for search algorithms in simulation experiments (and analytically when tractable). We reconsider the computer implementation of the Hoover–Perez search procedure, especially multiple search paths, the final-model selection criterion, orthogonalization, choice of significance levels, and their use of evidence on sub-sample significance.

Next, we discuss several key improvements to such algorithms. First, we advocate searching additional paths, including all initially-feasible search paths, as well as block tests, not just individual-coefficient tests. Secondly, we consider the potential value-added from pre-search simplifications based on sequential *F*-tests using non-stringent significance levels: variables found insignificant at this stage are excluded from further consideration, so reduce the size of the ‘general model’ from which path searches commence. Taken together, these stages could either increase or decrease the number of paths searched, depending on the properties of the model under analysis. Thirdly, we use parsimonious encompassing to discriminate between candidate models at the end of the first round of path searches: some models are dominated on this basis, and so are removed. If no unique model results, the union of the contenders constitutes the ‘general model’ for a fresh path-search iteration, and so on. When a unique model results from such stages, the algorithm terminates; otherwise it proceeds to select one model by an information criterion (rather than best fit, as in Hoover and Perez). Finally, the sub-sample significance of the coefficients in that model is evaluated by their proposal to examine overlapping split-samples, potentially assigning a reliability from zero to 100%.

The resulting algorithm (called *PcGets*) is then applied to the experiments in Lovell (1983) to evaluate the resulting improvement (or deterioration), and we report the outcomes in the four experiments where the procedure in Hoover and Perez performed least well. Notable improvements are found, revealing that their findings – in themselves a great improvement on earlier unstructured searches – are a lower bound on how well the methodology can perform when fully implemented. Finally, we apply *PcGets* to the general model proposed for US M1 by Baba, Hendry, and Starr (1992) to see whether it can perform well as an empirical tool – where the data-generation process is not a special case of the general model – and show that it selects a final parsimonious model close to that reported by those authors.

2. THE EXPERIMENTAL DESIGN

Lovell (1983) sought to evaluate the ability of some ‘model-selection methods’ to locate a single conditional equation (with 0 to 5 regressors) from a large macroeconomic database (containing up to 40 variables, including lags). In practice, he found none of the methods worked well, the legacy of which seems to be a professional belief that modelling *per se* (called ‘data mining’ in some circles) is bad. Although *Gets* is a more structured approach than any considered by Lovell, requiring an algorithm to select the correct sub-set of causal variables from 40 regressors is demanding. Hoover and Perez allow for near misses, as well as perfect matches, but we propose an alternative ‘performance measure’ in section 3.

Hoover and Perez extend the Lovell database, since it is a ‘neutral testbed’, and a ‘realistic’ choice, implicitly embodying many dozens of relations between variables – as in real economies – and of a scale and complexity that can occur in macroeconomics. Conversely, the experiment is ‘unrealistic’ in that the regressors are fixed in repeated samples, the parameters constant, and

the data-generation process (DGP) is a special case of the initial model, with homoscedastic, normal, innovation errors. Consequently, search procedures that conduct mis-specification tests will perform worse than those that do not, despite the obvious empirical drawbacks of the latter.

Their experiments generated a simulation outcome for one variable (y_t) as a linear function of the observed values ($x_{i,t}$, $i = 1, \dots, k$) of a subset of other database variables with parameter vector β , plus a set of random numbers $\epsilon_t \sim \text{IN}[0, \sigma_\epsilon^2]$:

$$y_t = \sum_{i=1}^k \beta_i x_{i,t} + u_t \quad \text{with} \quad u_t = \rho u_{t-1} + \epsilon_t. \quad (1)$$

Then y_t is regressed on all the database variables and their lags as in:

$$y_t = \sum_{j=1}^4 \alpha_j y_{t-j} + \sum_{i=1}^{18} \sum_{j=0}^1 \gamma_{i,j} x_{i,t-j} + \omega_t. \quad (2)$$

Consequently, (2) is the general unrestricted model (GUM), and the $\{x_{i,t}\}$ are strongly exogenous for the $\{\gamma_{i,0}\}$ (see Engle, Hendry, and Richard (1983)).

To ensure ‘stationary’ regressors, Hoover and Perez tested the original levels of the variables for unit-roots, and differenced each data series till that hypothesis could be rejected. Unfortunately, their objective was not fully achieved as figure 1 shows. The eight panels record the actual time series of the 18 regressors, organized in blocks of increasing variance, and the observed regressand (in the lower-right panel), numbered as in Hoover–Perez Table 1. Not only are the scales and ranges hugely different between variables, the second moments are non-constant over time, particularly for the dependent variable. Fortunately, because strong exogeneity holds in (2), such ‘mis-behaviour’ may be of little consequence, as can be checked both analytically (see e.g. Hendry (1995, ch. 4)), and by re-running their experiments using differences of the logs of the variables (see, e.g., the final column in Table 2 below).

3. COSTS OF INFERENCE AND COSTS OF SEARCH

We prefer a different metric to judge ‘success’, distinguishing between the costs of inference *per se*, and those of search. Statistical tests have non-degenerate null distributions (non-zero size) and usually non-unit power. Consequently, even if the DGP were derived *a priori* from economic theory, when an investigator does not *know* that the resulting model is ‘true’, but seeks to test conventional null hypotheses on its coefficients, then inferential mistakes will occur in general. In the present setting, the DGP is (1). Hypothesis tests of the form $H_0: \beta_i = 0$ will reject with a probability dependent on the non-centrality parameter of the test. Given the regression context, we consider t-tests, or more precisely, their squares, denoted $t^2(n, \psi^2)$ for n degrees of freedom, where ψ^2 is the non-centrality parameter.

For large n , $t^2(n, \psi^2)$ is distributed as a non-central $\chi^2(1, \psi^2)$, which in turn is approximately $h\chi^2(m, 0)$, where h and m are determined by ψ^2 to match the first two moments of the central and non-central χ^2 s (see e.g., Hendry (1995, ch. 13)):

$$h = 1 + \frac{\psi^2}{1 + \psi^2} \quad \text{and} \quad m = 1 + \frac{\psi^4}{1 + 2\psi^2},$$

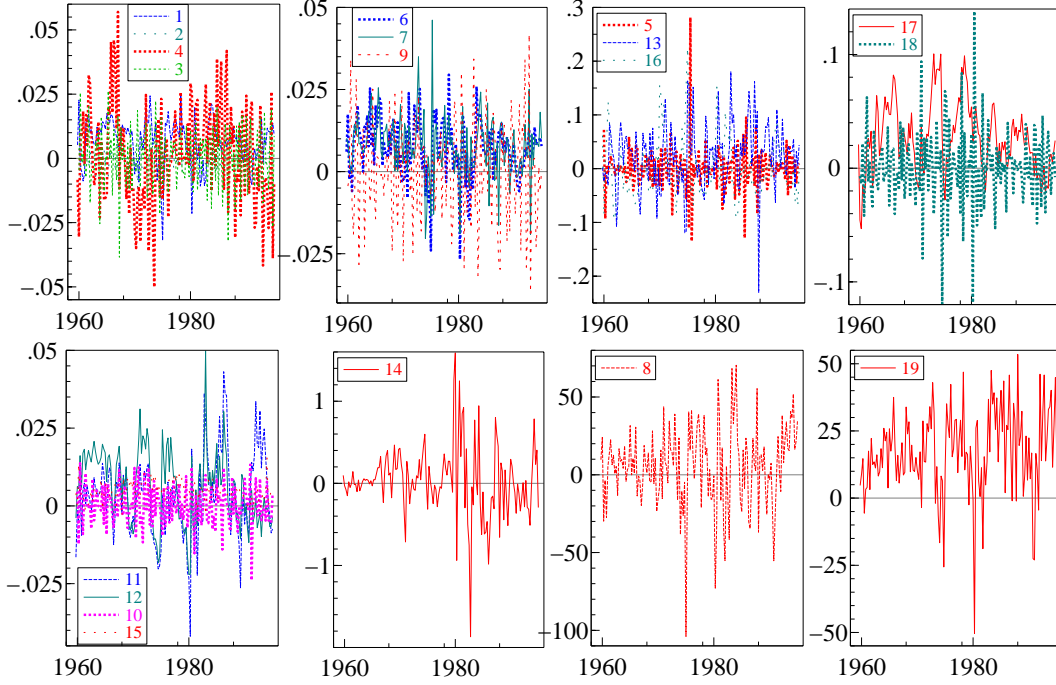


Figure 1. Graphs of the Hoover and Perez time-series data

and for a critical value c_α :

$$\Pr(\chi^2(1, \psi^2) \geq c_\alpha) \simeq \Pr(h\chi^2(m, 0) \geq c_\alpha) = \Pr(\chi^2(m, 0) \geq \frac{c_\alpha}{h}).$$

Since $E[\chi^2(1, \psi^2)] = 1 + \psi^2$, we use $\psi^2 = 4$ to represent a theoretical t -value of 2 (corresponding to $\Pr(|t| \geq 2|H_0) \simeq 0.05$), so $h = 1.8$ and $m = 2.8$. Approximating by $m = 3$ delivers $\Pr(\chi^2(3, 0) \geq 2.22) \simeq 0.5$ for $c_\alpha = 4$. Thus, even in a large sample, there is only a 50-50 chance of retaining a variable with a theoretical t^2 of 4 when the critical value is also 4. Raising c_α to 6.6 (noting $\Pr(|t| \geq 2.575|H_0) \simeq 0.01$) produces $\Pr(\chi^2(3, 0) \geq 3.67) \simeq 0.3$. When $\psi^2 = 9$, then $h = 1.9$ and $m \simeq 5$, so $\Pr(\chi^2(5, 0) \geq 3.47) \simeq 0.63$ for $c_\alpha = 6.6$, and the power of detection has risen sharply, but will still lead to more than 35% mis-classifications, although the effect is *only tested once*. Finally, for $\psi^2 = 16$, then $h \simeq 2$ and $m \simeq 9$, implying $\Pr(\chi^2(9, 0) \geq 3.3) > 0.95$ when $c_\alpha = 6.6$, so such a variable will almost always be retained.

Despite ‘knowing’ (1), low signal-noise variables will rarely be retained when tested, so the DGP would not be selected very often. Such a setting corresponds to a ‘pre-test’ problem: testing the truth will lead to false rejections (see e.g., Bock, Yancey, and Judge (1973), and Judge and Bock (1978)). Let p_i^{dgp} denote the probability of retaining the i^{th} variable when commencing from the DGP using a selection-test significance level α . Then $1 - p_i^{dgp}$ is the cost of inference. This is distinct from the costs of searching a large database for the DGP, where potentially an investigator might retain non-causal variables as well as drop DGP regressors more frequently. Let p_i^{gum} denote the probability of retaining the i^{th} variable when commencing from the GUM, also using significance level α . We measure search costs by $p_i^{dgp} - p_i^{gum}$, namely in relation

to the corresponding probabilities of retaining variables when commencing from the DGP. For irrelevant variables, $p_i^{dgp} \equiv 0$, so the whole cost of retaining adventitiously-significant variables is attributed to search, plus any additional costs for failing to retain relevant variables. The former can be lowered by increasing the required significance levels of selection tests, but at the cost of reducing the latter. However, it is feasible to lower the former and raise the latter simultaneously by an improved search algorithm, subject to the bound of attaining the same performance as knowing the DGP from the outset.

4. COMPUTER IMPLEMENTATION IN HOOVER AND PEREZ

To implement the *Gets* approach in a computer algorithm, Hoover and Perez ‘mechanize’ all decisions, and doing so throws light on several key methodological and practical modelling issues. Starting from the GUM, they first apply the battery of mis-specification tests to check its congruence. If that is acceptable, the algorithm follows a number of search paths, first deleting a variable which satisfies the reduction criterion – subject to the diagnostic tests remaining insignificant – and repeating that sequence till a violation of congruence occurs, when the algorithm terminates. Multiple paths can lead to multiple models, so after all paths are explored, Hoover and Perez select the model that fits best. If the GUM is unacceptable, either the ‘offending’ test is removed from the battery applied at later stages, or if more than one test fails, the experimental trial is discarded and a new sample drawn.

4.1. Multiple search paths

The most important of their operational steps is that of following a number of reduction search paths, each terminated either by no further feasible reductions, or diagnostic test rejections. Consequently, their algorithm avoids getting ‘stuck’ in a search that initially inadvertently deletes a DGP variable, and retains many other variables as proxies. This ‘multiple path’ implementation is to be welcomed, and opens up a range of potentially important improvements to modelling practice (see Mizon (1977) for an earlier advocacy of that approach). Potential improvements include trying all single-variable deletion starting points, and feasible block deletions. Theoretically, searching only one path is worse than their choice of 10, or all initially-insignificant variables, but the ‘optimal’ number of paths to search remains an open question. Although a large number of additional selection and diagnostic tests is calculated along these path searches, nevertheless overall *better size properties result*.

4.2. Final selection

Secondly, if no unique choice of model has occurred after searching all 10 paths, Hoover and Perez select the model with the smallest equation standard error. Such a rule tends to select models which systematically overfit, since the minimum equation standard error occurs for $|t| > 1$. Hence, they retain an excess number of variables at any given nominal selection-test size, and understate how well a *Gets* approach can do. Indeed, such findings are predictable from the ‘LSE’ methodology, and a strict application of its principles (via encompassing tests) should be able to significantly reduce the problem of overfitting, as we discuss below. Improvements are feasible here even without the additional search steps we propose below.

4.3. Simulating joint selection and diagnostic testing

The behaviour of jointly selecting variables with diagnostic testing has eluded theoretical analysis, and although Monte Carlo is always problem dependent, the Hoover–Perez findings cohere with what theory exists. Since the null is true, any diagnostic testing acts as a constraint on reducing the GUM, so will increase the overall size of the selection process (in terms of retaining spurious variables). Thus, the choices of which – and how many – diagnostic tests to compute, are important if a search is to progress towards simplification. Moreover the tests must have the correct size, as well as meet a stringent significance level (e.g., 1% when tests for residual autocorrelation, non-normality, non-constancy, and heteroscedasticity are all used). Conversely, by checking diagnostically at every stage, false deletions may be detected, leading to a final selection closer to the DGP despite considerable additional testing, refuting any *a priori* belief that additional testing must increase overall size.

Ensuring that reductions are congruent has long been a key plank of the ‘LSE’ approach (see e.g., Hendry (1980)). For non-stationary time series, a search algorithm with diagnostic testing could improve over simply testing the DGP for retaining variables, since there will be cases where candidate variables do not satisfy a selection criterion, yet their deletion induces a significant diagnostic. Conversely, by constraining all deletions by congruence, ‘spurious’ effects may be retained more often. Combining these ideas suggests using a tighter significance level than the conventional 5% for both forms of test.

The simulations in Hoover and Perez could either understate or overstate how well *Gets* might do in practice. On the former, when a diagnostic test of the DGP is significant (as must happen by chance with tests of non-zero size), it is dropped from the checking set. Consequently, an ever-increasing problem of that type could lurk undetected, possibly distorting inference through inappropriate estimated parameter standard errors. It seems better to initially increase the nominal rejection level, and if that higher level is exceeded during any search path, then stop the search; tests that are significant in the GUM can sometimes cease to be significant as reduction proceeds, but sometimes increase to reveal a flawed path.

On the latter (overstating the performance of *Gets*), more research is needed on cases where the DGP is rejected against the GUM. Hoover and Perez use a ‘2-test rejection criterion’ to discard the generated data sample, which probably favours *Gets*, since reduction would be problematic with such aberrant data. With 1% significance levels, such cases will rarely arise under the null, but we are close to metaphysics here: does an algorithm fail by ‘overfitting’ on such unlikely samples, or would investigators of a non-replicable world conclude that such features really were aspects of the DGP? Indeed, how could any empirical investigator ‘know’ that the economy was simpler empirically than the data suggest? An approach of fitting the ‘true’ equation, then diagnostically testing it against a range of alternatives need not fare any better in such a case, unless the investigator knew the truth, and knew that she knew it, so *no* tests were needed. Since we view research as embedded in a progressive strategy, the ‘problem’ will be clarified as additional data accrue, since these are likely to be more ‘representative’.

4.4. Orthogonalization

Next, the role of variable orthogonalization is highlighted by their study. The use of differences to ‘achieve stationarity’ defines the ‘mine’ to be explored, and as their Table 2 shows, induces fairly low correlations between the potential regressors, other than their own lags. Hoover and

Perez find the second lagged-dependent variable, the least orthogonal to the first, to be most often incorrectly selected, even in addition to the first: such a finding poses a problem for the underlying statistical theory. Empirically, in a second-order autoregression for GCQ , the coefficient of GCQ_{t-2} is -0.17 ($t = -2.1$), but simulation of the t-value on the second lag in a stationary scalar autoregression – when only the first lag matters and from zero to 20 extraneous variables are included – revealed no substantial ‘over-sized’ outcomes. Thus, an interaction with selection seems to occur: future research should ascertain whether this is an ‘over-estimation’ of dynamics, or merely ‘spreads’ the lag structure across the two lags.

4.5. Evaluation

Hoover and Perez define various categories of ‘success’ and ‘failure’. This is a useful refinement, and helps reveal against which alternatives ‘incorrect’ choices are made. Some investigators might regard category 2 as the worst kind of failure, by spurious overfitting; whereas others would condemn the inability of the search to detect effects that might be of policy significance. Hoover and Perez also carefully refer to the ‘algorithm’ failing (rather than the methodology), but towards the end of their paper, that distinction becomes blurred: however, improvements in the algorithm could alter the simulation outcomes without changing the methodology, and does so, as we show below. Section 3 explained why we prefer a different evaluation standard, although in our replication of their experiments, we retain the one they propose.

4.6. Significance levels

It is worth distinguishing between the significance levels of the selection tests and those of the diagnostic tests. Lowering the significance level of the diagnostic tests from (say) 0.05 to 0.01 reduces the overall size (this is the difference in powering up 0.95 and 0.99 repeatedly) so few chance ‘multiple rejects’ occur, without greatly affecting the overall power of the procedure. Changing the significance level of the selection t-tests also reduces the empirical size, but lowers the power more noticeably for variables with theoretical t-values of around 2 or 3. This trade-off is therefore selectable by an investigator. Consequently, whether or not the ‘LSE’ approach over or under selects is not intrinsic to it, but to how it is used. Smaller significance levels (1% versus 5%) have much to commend them, especially given the support from the shape of the size-power trade-off that Hoover and Perez report. We corroborate their results here, and find that 1% for all tests at the sample sizes current in macro does well, and dominates 5% dramatically on overall size, without much power loss once $|t| > 2.5$.

4.7. Sub-sample testing

Finally, their ‘cross-validation’ approach merits further exploration. Since a central t-test has a mean of zero but wanders around the origin, the probability is low that an effect which is significant only by chance in the full sample will also be significant in two sub-samples. Conversely, a non-central t diverges, so should be significant in both sub-samples (perhaps at a lower significance level since the samples are smaller). Certainly, this idea works well in their Monte Carlo, and reduces ‘overfitting’ relative to their original algorithm. We believe that it will be a powerful

strategy for model selection when breaks occur in some of the marginal relations over one of the sub-samples, and is used in a modified form in our program, *PcGets*, as we now discuss.

5. AN IMPROVED ALGORITHM

The preceding discussion noted several potential improvements to the algorithm in Hoover and Perez. This section describes how these are implemented in *PcGets*.¹

5.1. Pre-search reductions

First, groups of variables are tested in the order of their absolute t-values, commencing with a block where all the p-values exceed 0.9, and continuing down towards the pre-assigned selection criterion, when deletion must become inadmissible. Consequently, these ‘pre-selection’ reduction tests sequentially increase the number of variables tested from the least significant to the most, terminating when the first rejection occurs. A less stringent significance level is used at this step, usually 10%, since the insignificant variables are deleted permanently, with a new GUM being formed from the remainder. The sequential ordering of the tests and the lower significance level help protect against missing important variables which are ‘hidden’ in a morass of irrelevance, while detecting cases where almost no variables matter and thereby avoiding over-fitting. If no test is significant, the F-test on all variables in the GUM has been calculated, establishing that there is nothing to model. This step therefore helps most when no, or few, variables matter, and can greatly reduce the number of paths that need to be searched.

5.2. Additional paths

Secondly, blocks of variables constitute feasible search paths, in addition to individual-coefficients. These additional tests operate like the block F-tests in the preceding sub-section but along search paths. Then all paths that also commence with an insignificant t-deletion are explored. These two steps obviously increase the number of paths searched.

5.3. Encompassing

A third improvement is to use encompassing tests between the candidate congruent selections at the end of path searches. Thus, *PcGets* uses parsimonious-encompassing tests to select between the models at stage H in Hoover and Perez when no unique choice has resulted from the first-round multiple-path searches. Each contender is tested against their union, dropping those which are dominated by, and do not dominate, another contender. If a unique model results, select that; otherwise, if some are rejected, form the union of the remaining models, and repeat this round till no encompassing reductions result. That union then constitutes a new starting point, and the complete path-search algorithm repeats till the union is unchanged between successive rounds.

¹*PcGets* is an Ox Package (see Doornik (1998) and Hendry and Krolzig (1999a)) designed for general-to-specific modelling, presently focusing on reduction approaches for linear, dynamic, regression models.

The simulation results suggest that such encompassing checks help ‘control’ the overall size of the path-search algorithm: different paths can keep different relevant and irrelevant variables, and the encompassing ‘sieve’ then eliminates most of the irrelevant ones against a union which retains most of the relevant.

5.4. Information criteria

When a union coincides with the original GUM, or with a previous union, so no further feasible reductions can be found, *PcGets* selects a model by an information criterion. The three ‘final-selection’ rules presently computed are AIC (Akaike information criterion: see Akaike (1985)), SC (Schwarz criterion, also called BIC: see Schwarz (1978)) and HQ (Hannan–Quinn: see Hannan and Quinn (1979)), defined as:

$$\begin{aligned} AIC &= -2 \log L/T + 2p/T, \\ SC &= -2 \log L/T + p \log(T)/T, \\ HQ &= -2 \log L/T + 2p \log(\log(T))/T, \end{aligned}$$

where L is the maximized likelihood, p is the number of parameters and T is the sample size. SC has the best operational characteristics in our own Monte Carlo experiments when no unique candidate model was selected by the earlier steps. For $T = 140$ and $p = 40$, minimum SC corresponds approximately to the marginal regressor satisfying $|t| \geq 1.9$.

5.5. Sub-sample reliability

Next, for that finally-selected model, the sub-sample reliability of its coefficients is evaluated by the proposal in Hoover and Perez for overlapping split-sample tests. Thus, *PcGets* concludes that some variables are definitely excluded (i.e., not in the final model); some definitely included (i.e., they satisfy the selection criteria in the final model and in both sub-samples), and some have an uncertain role, varying from a reliability of 25% (included in the final model, but insignificant – probably acting as a proxy or because dropping induces a significant diagnostic – and insignificant in both sub-samples), through to 75% (significant overall and in one sub-sample, or in both sub-samples).

5.6. Significant mis-specification tests

Finally, if the initial mis-specification tests are significant at the pre-specified level in the Monte Carlo, we retain that sample, but raise the required significance levels, terminating search paths only when these higher levels are violated. Computed coefficient standard errors might be biased relative to sampling standard deviations in such a case (although the actual errors are normal white noise in the experiments), but this approach seems preferable to omitting the test altogether, or re-sampling. Empirical investigators would undoubtedly re-specify the GUM on rejection, so we may slightly favour *Gets* by this strategy (with four tests at the 1% level, the probability that one will reject in any replication is about 4%, and is negligible that two will).

Table 1. COMFAC and ADL representations of selected Hoover–Perez DGPs

1	WN	$y_t = 130\varepsilon_t, \quad \varepsilon_t \sim \text{IN}(0, 1)$
2	COMFAC ADL(1,1)	$y_t = 130u_t, \quad u_t = 0.75u_{t-1} + \varepsilon_t, \quad \varepsilon_t \sim \text{IN}(0, \frac{7}{16})$ $y_t = 0.75y_{t-1} + 130\varepsilon_t$
7	COMFAC ADL(1,1)	$y_t = 1.33\Delta FM1DQ_t + 9.73u_t, \quad u_t = 0.75u_{t-1} + \varepsilon_t, \quad \varepsilon_t \sim \text{IN}(0, \frac{7}{16})$ $y_t = 0.75y_{t-1} + 1.33\Delta FM1DQ_t - 0.975\Delta FM1DQ_{t-1} + 9.73\varepsilon_t$
9	COMFAC ADL(1,1)	$y_t = 0.67\Delta FM1DQ_t - 0.023\Delta^2 GGEQ_t + 4.92u_t, \quad u_t = 0.75u_{t-1} + \varepsilon_t, \quad \varepsilon_t \sim \text{IN}(0, \frac{7}{16})$ $y_t = 0.75y_{t-1} + 0.67\Delta FM1DQ_t - 0.5025\Delta FM1DQ_{t-1} - 0.023\Delta^2 GGEQ_t$ $+ 0.01725\Delta^2 GGEQ_{t-1} + 4.92\varepsilon_t$

Legend: WN, COMFAC and ADL(1,1) respectively denote white noise, common-factor restrictions, and autoregressive-distributed lag with current and one lag in the dependent and regressor variables.

6. PCGETS IN ACTION

As just described, our program *PcGets* (see Krolzig and Hendry (1999)) allows as options block pre-search tests, user choice of all significance levels, final-model selection criteria, and percentage overlap in the sub-sample split; t and F-path searches and encompassing tests are always implemented. To examine its performance, we re-ran the experiments in Hoover and Perez (and our own Monte Carlo), but here we only record the outcomes for the four experiments where they found *Gets* to have the worst performance. These are shown in table 1. The corresponding outcomes are shown in table 2, where pre-search reduction tests denote the sequential F-tests discussed in § 5.1.²

Comparing these results with the corresponding findings in Tables 4 or 7 in Hoover and Perez, the DGP is found in:

- HP1 97.2% of the cases [at 1% size], against 79.9% in their table 7;
- HP2 60.2% of the cases [at 1% size], against 0.8% in their table 7;
- HP2 improved to 85.2% of the cases [at 1% size] if pre-search reduction is used;
- HP2 8.4% of the cases [at 5% size], against 0% in their table 4 –
- HP2 improved to 76.9% of the cases [at 5% size] if pre-search reduction is used;
- HP7 59.0% of the cases [at 1% size], against 24.6% in their table 7.

We deem these to be substantial improvements, even though the settings in the present *PcGets* algorithm are almost certainly not optimal, so superior performance to that reported in table 2 remains quite feasible. In re-running these experiments, the probability of finding the truth is between 97.2% and 0.0% – depending on the specification of the DGP. Instead of focusing on whether the DGP has been found or not, we would focus on the statistical qualities of the selected model relative to the DGP. As discussed above, the nature of the DGP might make it impossible to find it on the selection criteria adopted.

²The GUM is equation (2) where y is ΔGCQ , and $\{x_j\}$ denotes their variables: $\{\Delta DCOINC, \Delta^2 GD, \Delta^2 GGEQ, \Delta GFEQ, \Delta^2 GGFR, \Delta GNPQ, \Delta GYDQ, \Delta GPIQ, \Delta^2 FMRR, \Delta^2 FMBASE, \Delta FM1DQ, \Delta FM2DQ, \Delta FSDJ, \Delta FYAAC, \Delta LHC, \Delta LHUR, \Delta MU, \Delta^2 MO\}$.

Table 2. Simulation results with *PcGets*

Experiment	HP1	HP2	HP2	HP2	HP2	HP7	HP7	HP9	HP9	HP9	HPL2
Significance level	0.01	0.01	0.01	0.05	0.05	0.01	0.05	0.01	0.01	0.05	0.05
Pre-search reduction tests	—	—	yes	—	yes	—	—	—	yes	—	
DGP: <i>t</i>-values											
Δy_{t-1}		12.95	12.95	12.95	12.95	12.49	12.49	12.61	12.61	12.62	12.95
ΔFM1DQ_t						15.14	15.14	15.01	15.01	15.02	
$\Delta \text{FM1DQ}_{t-1}$						-8.16	-8.16	-8.47	-8.47	-8.47	
$\Delta^2 \text{GGEQ}_t$								-0.63	-0.63	-0.64	
$\Delta^2 \text{GGEQ}_{t-1}$								0.47	0.47	0.47	
Selection probabilities											
Δy_{t-1}		1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
Standard error		0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
ΔFM1DQ_t						1.0000	1.0000	1.0000	1.0000	1.0000	
Standard error						0.0000	0.0000	0.0000	0.0000	0.0000	
$\Delta \text{FM1DQ}_{t-1}$						0.9970	0.9980	0.9960	0.9960	0.9980	
Standard error						0.0017	0.0014	0.0020	0.0020	0.0014	
$\Delta^2 \text{GGEQ}_t$								0.0560	0.0160	0.1680	
Standard error								0.0073	0.0040	0.0118	
$\Delta^2 \text{GGEQ}_{t-1}$								0.0430	0.0180	0.1460	
Standard error								0.0064	0.0042	0.0112	
Non-deletion probability											
Sample mean	0.0019	0.0242	0.0088	0.1072	0.0199	0.0243	0.1017	0.0246	0.0088	0.1045	0.1108
Standard error	0.0002	0.0008	0.0005	0.0015	0.0007	0.0008	0.0015	0.0008	0.0005	0.0016	0.0016
Selected model											
DGP found	0.9720	0.6020	0.8520	0.0840	0.7690	0.5900	0.1050	0.0000	0.0000	0.0000	0.0870
Non-DGP var. included	0.0280	0.3980	0.1480	0.9160	0.2310	0.4100	0.8950	0.4200	0.1480	0.8930	0.9130
DGP var. not included	0.0000	0.0000	0.0000	0.0000	0.0000	0.0030	0.0020	0.9970	0.9990	0.9890	0.0000
DGP is dominated	0.0260	0.3830	0.1030	0.9080	0.2100	0.3900	0.8900	0.9250	0.9180	0.9680	0.9030
Specific is dominated	0.0020	0.0150	0.0450	0.0080	0.0210	0.0200	0.0050	0.0010	0.0300	0.0000	0.0100

Based on $M = 1000$ replications of the DGP with sample size $T = 100$ and $\eta = 0.01$.

The HP1 experiment demonstrates the effect of the pre-search F-tests on limiting the ‘size’ of the model reduction process by testing the GUM against reductions up to an empty model.³ If the F-test that all coefficients of the model are zero cannot be rejected, then the empty model is reported as the valid reduction of the GUM. The resulting average non-deletion probability of 0.2% is much lower than the nominal size of 1% per variable, and dramatically lower than that found by Hoover and Perez. Such an outcome can arise because the F-test against the GUM would have a 1% size under the null, if it were the only test implemented, so would induce an average non-deletion probability of 0.025% times the number of variables retained when the null model is rejected. However, as the other experiments reveal, this size reduction is not at the cost of increased elimination of variables that actually matter.

The HP2 experiments if anything show an even more dramatic improvement: if our recommended 1% size is used, the DGP is recovered more than 60% of the time against their finding it less than 1% (at 1%); if pre-search tests are used, then we find the DGP more than 85% of the time. Thus, huge scope existed for improvements in the search algorithm; and may well still do relative to our present procedure.

Hoover and Perez evaluate the success or failure of their *Gets* algorithm on the basis of a comparison of the selected model and the autoregressive-distributed lag (ADL) transformation of the initial COMFAC representation of the DGPs. This is consistent with the theoretical assumption of *Gets* that the errors form a mean-innovation process, which excludes the COMFAC assumption of AR(1) errors. We follow them in evaluating the Monte Carlo experiments with *PcGets* using the ADL representation of the DGP – although the results of *PcGets* modelling are more impressive. In the worst-case scenario, HP9, the DGP involves the variables $\Delta^2 GGEQ_t$ and $\Delta^2 GGEQ_{t-1}$ whose t-values (evaluated in the true model) are on average -0.63 and 0.47 , respectively. So even if one started with the truth, based on statistical criteria $\Delta^2 GGEQ_t$ and $\Delta^2 GGEQ_{t-1}$ would usually be removed. In a world like HP9, a data-driven approach would rarely detect that $\Delta^2 GGEQ$ is a part of the DGP.

For these reasons, we prefer to check whether the deviation of the ‘specific’ model found by *PcGets* from the ‘true’ model nevertheless results in a sound model that, based on statistically criteria, could not have been improved by knowing the truth. We consider an encompassing test between the ‘true’ model and the ‘specific’ model found by *PcGets*. As long as *PcGets* is able to find a model that is not dominated by the ‘true’ model, the reduction process has been a success. If the specific model is dominated by the ‘true’ model, then the search algorithm has failed. Our results indicate that the risk of finding a model which is dominated by the DGP is extremely small: in most cases the risk is less than one percent and never greater than 4.5%. However, there might be a scope for further improvements by future developments.

6.1. Reducing heteroscedasticity

The graphs in figure 1 suggest that many of the data series are heteroscedastic, with an increasing variance over time. This is due to taking differences of the original variables, rather than their log transforms, when the scale of the US economy has grown substantially. Thus, we conducted several experiments on the log-differences to check that this feature had not induced artefacts into the simulation results. Since the GUM, DGP, and any selected model thereof were all valid

³*PcGets* pre-tests whether the mean is significantly different from zero, in which case a constant is introduced under the null.

conditional representations with strongly-exogenous regressors, we anticipated little change, although the heteroscedasticity test could have been a powerful help in retaining relevant variables as their deletion might have induced significant diagnostic outcomes. The final column of table 2 (denoted HPL2) records the outcomes in experiment HP2, and comparison with column 4 reveals closely similar results.

6.2. Integrated variables

To date, *PcGets* conducts all inferences as if they were $I(0)$. Given the results in Sims, Stock, and Watson (1990), most reduction tests will be valid even when the data are $I(1)$. Only a t- or F-test on a unit-root effect will require a non-standard critical value. For example, if all but one variable is transformed to $I(0)$, the last being $I(1)$, then invalid inference is bound to occur, using critical values that are too small. The empirical examples on $I(1)$ data provided below do not reveal problems, but in principle it would be useful to implement cointegration test reductions and appropriate transformations in a Stage 0, prior to Stage I reductions.

Wooldridge (1999) shows that the diagnostic tests on the GUM, and simplifications thereof, including those used here, remain valid for integrated time series. Thus, mis-specification inference should not be distorted, avoiding the difficulty of conditionally altering critical values as a function of ‘pre-tests’ for unit roots and cointegration.

7. BHS REVISITED

To examine the behaviour of the search algorithm on an empirical problem, we applied it to the model of US narrow-money demand in Baba, Hendry, and Starr (1992), denoted by the acronym BHS. The equation used as a baseline corresponds to their GUM, but expressed in terms of real M1 ($m - p$) and inflation (Δp), where lower case denotes logs.⁴ y is (log) GNP, R_l , R_1 , R_{ma} , R_{na} , and R_{sna} are yields on 20-year Treasury bonds, 1-month bills, M2 instruments, NOW and super-NOW accounts. Finally, V is a measure of the volatility of R_l , $SV = \max(0, S) * V$ where $S = R_l - R_1$, and $I1980$ is a dummy variable which is +1, -1 in 1980(2), 1980(3) reflecting changes in credit control (see appendix 10, and Baba, Hendry, and Starr (1992), p38, for details about the data series). The GUM is shown in table 3.

In table 4, t_{ur} is the *PcGive* test for cointegration (see Hendry and Doornik (1996, p235)), R^2 is the squared multiple correlation, $\hat{\sigma}$ is the residual standard error, and the diagnostic tests are of the form $F_j(k, T - l)$ which denotes an F-test against the alternative hypothesis j for: 4th-order serial correlation (F_{ar} : see Godfrey (1978)), 4th-order autoregressive conditional heteroscedasticity (F_{arch} : see Engle (1982)), heteroscedasticity (F_{het} : see White (1980)); the RESET test (F_{reset} : see Ramsey (1969)); and a chi-square test for normality ($\chi^2_{nd}(2)$: see Doornik and Hansen (1994)): * and ** denote significance at the 5% and 1% levels respectively, and na denotes not available. Jt and V are the variance-change and the joint parameter-constancy tests from Hansen (1992). For convenience, we report re-estimates of the preferred model (22) from BHS.

⁴A GUM based on nominal money was also used, but had little impact on the selected model, leading to a slightly more parsimonious specification with 12 regressors ($SC = -10.47$).

Table 3. BHS GUM: 1960(3)–1988(3)

Lag	0	1	2	3	4	5	6	Sum
$m - p$	−1	1.070	−0.402	0.003	−0.140	0.423	−0.167	−0.210
	−	0.099	0.134	0.118	0.106	0.103	0.072	0.030
y	0.196	−0.015	−0.050	−0.088	0.111	−0.066	0.028	0.117
	0.061	0.077	0.072	0.069	0.069	0.068	0.053	0.014
Δp	−0.699	0.161	−0.350	−0.288	−0.279	−0.002	−	−1.460
	0.119	0.141	0.132	0.136	0.132	0.138	−	0.296
V	0.595	0.156	−	−	−	−	−	0.751
	0.419	0.418	−	−	−	−	−	0.117
R_t	−0.690	−0.578	−	−	−	−	−	−1.270
	0.228	0.291	−	−	−	−	−	0.256
R_1	0.219	0.137	−	−	−	−	−	0.356
	0.189	0.216	−	−	−	−	−	0.200
R_{ma}	−0.249	0.291	−	−	−	−	−	0.042
	0.096	0.093	−	−	−	−	−	0.078
R_{na}	1.010	−1.610	0.885	−	−	−	−	0.279
	0.528	0.801	0.432	−	−	−	−	0.106
R_{sna}	−0.720	1.660	−0.867	−	−	−	−	0.075
	0.736	1.260	0.647	−	−	−	−	0.121
$I1980$	0.013	−	−	−	−	−	−	0.013
	0.004	−	−	−	−	−	−	0.004
SV	1.390	9.230	−12.800	−	−	−	−	−2.220
	5.380	6.330	2.590	−	−	−	−	7.210
1	0.292	−	−	−	−	−	−	0.292
	0.044	−	−	−	−	−	−	0.044

$$\begin{aligned}
\Delta \left(\widehat{m-p} \right)_t &= \underset{(0.021)}{0.358} + \underset{(0.003)}{0.013 I1980_t} + \underset{(0.070)}{0.370 \Delta A y_t} - \underset{(0.129)}{1.066 \Delta_4 p_{t-1}} \\
&- \underset{(0.046)}{0.341 \Delta \hat{p}_t} - \underset{(0.049)}{0.260 \Delta R_{ma,t}} - \underset{(0.105)}{1.428 AS_t^*} - \underset{(0.063)}{0.985 AR_{1,t}} + \underset{(0.051)}{0.465 R_{nsa,t}} \\
&- \underset{(0.015)}{0.253 \left(m-p - \frac{1}{2}y \right)_{t-2}} - \underset{(0.098)}{0.348 \Delta_4 (m-p)_{t-1}} - \underset{(0.040)}{0.148 \Delta^2 (m-p)_{t-4}}
\end{aligned} \tag{3}$$

$$R^2 = 0.889, \hat{\sigma} = 0.391\%, F(11, 101) = 73.32, DW = 1.79, SC = -10.70$$

In contrast to the equilibrium-correction model in (3), the following analysis with *PcGets* is applied to the levels (as shown in table 3). After pre-selection F-tests at 10%, 11 paths were explored (at 1%), leading to two mutually-encompassing contenders. The union was formed, and 4 paths explored, leading back to the same two models, which only differed by the former

Table 4. BHS GUM diagnostics

$t_{ur} = -7.06^{**}$	$R^2 = 0.9984$	$\hat{\sigma} = 0.398\%$	$SC = -9.87$
$\chi_{nd}^2(2) = 0.22$	$F_{ar}(4, 71) = 0.82$	$F_{arch}(4, 67) = 0.61$	$F_{reset}(1, 74) = 0.39$
$Jt = 5.21$	$V = 0.12$	$F_{het}(., .)(na)$	$F_{Chow}(12, 63) = 0.84$

having Δp_{t-3} and the latter $(m - p)_{t-4}$. The encompassing tests yielded (probabilities shown in brackets):

$$M_1 : F(1, 96) = 4.76 [0.032]$$

$$M_2 : F(1, 96) = 6.50 [0.012]$$

Neither model is strongly dominated by the other; both are marginally acceptable reductions of their union. The first model was selected as dominant on all three information criterion, and is shown in (4): given the GUM, the program takes only a few seconds to select this equation.

$$\begin{aligned}
 (\widehat{m - p})_t = & \underset{(19.96)}{1.115} (m - p)_{t-1} - \underset{(6.79)}{0.419} (m - p)_{t-2} + \underset{(4.27)}{0.260} (m - p)_{t-5} \\
 & - \underset{(2.56)}{0.134} (m - p)_{t-6} + \underset{(13.70)}{0.113} y_t - \underset{(8.08)}{0.845} \Delta p_t - \underset{(3.63)}{0.408} \Delta p_{t-2} - \underset{(2.93)}{0.336} \Delta p_{t-3} \\
 & + \underset{(8.79)}{0.701} V_t - \underset{(11.62)}{0.751} R_{l,t} + \underset{(3.09)}{1.187} R_{na,t} - \underset{(3.09)}{1.857} R_{na,t-1} + \underset{(2.59)}{0.902} R_{na,t-2} \\
 & - \underset{(7.64)}{13.03} SV_{t-2} + \underset{(10.97)}{0.24} + \underset{(5.26)}{0.017} I1980_t \\
 & R^2 = 0.9976, \hat{\sigma} = 0.43\%, F(15, 97) = 2725, SC = -10.37
 \end{aligned} \tag{4}$$

't'-values are shown in parentheses (as that was the selection criterion), estimation was by OLS, and the full sample was used. Eleven coefficients were reported as 100% reliable, four as 75% (Δp_{t-3} , $R_{na,t}$, $R_{na,t-1}$ and $R_{na,t-2}$) and the remaining coefficient ($(m - p)_{t-6}$) as 50%. Note that R_{na} is zero till 1981, so the lack of sub-sample reliability in those coefficients is unsurprising.

The resulting long-run solution (cointegrating vector) is:

$$\begin{aligned}
 (m - p) = & \underset{(0.017)}{1.349} + \underset{(0.039)}{0.634} y - \underset{(0.22)}{4.22} R_l + \underset{(0.35)}{3.93} V + \underset{(0.38)}{1.31} R_{na} \\
 & - \underset{(1.15)}{8.92} \Delta p - \underset{(10.1)}{73.3} SV
 \end{aligned} \tag{5}$$

Once the specification in (4) is re-arranged in differences and the cointegrating vector from 5, we obtain an $I(0)$ equation close to BHS, with R_l and R_{na} instead of R_1 and R_{sna} .

Using no pre-selection tests, (i.e., only F and t-tests at 1%,) 27 paths were selected. The search set yielded six models of which 4 were dominated (and so excluded), leaving only one parsimonious-encompassing model for the remaining union, which became the unique reduction

with 18 variables, most in common with (4). No diagnostic test was significant in any selection, and 15 coefficients were 100% reliable, with the remainder 75%.

Finally, tightening the significance level on the pre-search F-tests from 10% to 1% leads to a simpler model with 13 variables, which omits $(m - p)_{t-6}$, $R_{na,t-1}$, and $R_{na,t-2}$. All these outcomes are as anticipated in terms of degree of parsimony, and lead to similar final selections.

8. CONCLUSION

Overall, it is many years since we have read a paper that has stimulated so many exciting new ideas and potential improvements. Hoover and Perez have demonstrated that the first mechanized general-to-specific algorithm performs dramatically better than we could have hoped: improvements in its design can only induce better behaviour on this type of problem, as *PcGets* has already shown. Enforcing congruence entails searching in the correct model class and excludes many non-viable contenders; multiple search paths avoid cul-de-sacs in the search space; encompassing eliminates potential candidate models; pre-search can focus the initial GUM without much risk of omitting key variables; information criteria improve selection over ‘best-fit’ rules; and split-sample analysis helps reveal adventitiously-significant effects. Replicating the analysis of the Lovell database for the experiments in Hoover and Perez by implementing these developments in *PcGets* shows that major improvements are feasible, and we suspect that further enhancements will appear.

The specification of the general unrestricted model remains central to the performance of the search process, and requires congruence. The *larger* the initial regressor set, the more likely adventitious effects will be retained. This suggests a central role for economic theory in ‘prior simplification’. But the *smaller* the GUM, the more likely key variables will be omitted. Thus, the reasonably low costs of search found by Hoover and Perez, and corroborated here, suggest adopting relatively generous parameterizations in the GUM.

Similarly, the less orthogonality, the more ‘confusion’ the algorithm faces, which can lead to a proliferation of mutually-encompassing models, where the final choice may only differ marginally (e.g., lags 1 and 2 versus 1 and 3). Careful prior analysis therefore remains essential, concerning the parameterization, functional form, choice of variables, lag length, any necessary indicators (including seasonals) and so. Transforming to $I(0)$ representations may deliver better empirical performance, but requires simulation analysis.

Choice of significance levels and search rules (block or individual) can substantively alter the performance of *Gets*, and we favour 1% as a ‘baseline’ for the types of problem considered by Hoover and Perez, and above. Similarly, it is crucial that the diagnostic tests be well behaved, comprehensive, and not too numerous. More research is needed on the role of the diagnostic tests under the alternative, not as false constructivism, but when they are the correct test for the problem at hand, and reveal potentially-serious mis-specifications.

8.1. The way ahead

On the operational front, we see many possible developments, including: forced search paths to test rival models; forced inclusions of variables deemed essential by an investigator; vector, and instrumental-variable, generalizations; adaptive critical values relative to the size of intermediate

models; an extended diagnostic test battery; and greater use of recursive statistics, *inter alia* especially in the context of a progressive research strategy.

On the analytic front, a more complete theory of the selection process now appears feasible, and is under development, including analyses of the role of pre-selection tests; the number of paths to search; the impact of the encompassing evaluation stage; the effect of diagnostic tests on selection probabilities; the role of structural breaks in marginal processes; and the control of adventitious selection. Some of our findings, together with further Monte Carlo evidence, are reported in Hendry and Krolzig (1999b) and Krolzig and Hendry (1999).

9. ACKNOWLEDGEMENTS

All the computations reported in this paper were carried out with the PcGets class in Ox. We are pleased to acknowledge financial support from the UK Economic and Social Research Council under grant L116251015, and are grateful to Mike Clements, Jurgen Doornik, Neil Ericsson, David Firth, Jon Faust, Grayham Mizon, and Neil Shephard for helpful comments.

REFERENCES

- Akaike, H. (1985). Prediction and entropy. In A. C. Atkinson and S. E. Fienberg (Eds.), *A Celebration of Statistics*, pp. 1–24. New York: Springer-Verlag.
- Baba, Y., D. F. Hendry, and R. M. Starr (1992). The demand for M1 in the U.S.A., 1960–1988. *Review of Economic Studies* 59, 25–61.
- Bock, M. E., T. A. Yancey, and G. C. Judge (1973). Statistical consequences of preliminary test estimators in regression. *Journal of the American Statistical Association* 68, 109–116.
- Doornik, J. A. (1998). *Object-Oriented Matrix Programming using Ox 2.0*. London: Timberlake Consultants Press.
- Doornik, J. A. and H. Hansen (1994). A practical test for univariate and multivariate normality. Discussion paper, Nuffield College.
- Engle, R. F. (1982). Autoregressive conditional heteroscedasticity, with estimates of the variance of United Kingdom inflations. *Econometrica* 50, 987–1007.
- Engle, R. F., D. F. Hendry, and J.-F. Richard (1983). Exogeneity. *Econometrica* 51, 277–304. Reprinted in Hendry, D. F., *Econometrics: Alchemy or Science?* Oxford: Blackwell Publishers, 1993; and in Ericsson, N. R. and Irons, J. S. (eds.) *Testing Exogeneity*, Oxford: Oxford University Press, 1994.
- Godfrey, L. G. (1978). Testing for higher order serial correlation in regression equations when the regressors include lagged dependent variables. *Econometrica* 46, 1303–1313.
- Hannan, E. J. and B. G. Quinn (1979). The determination of the order of an autoregression. *Journal of the Royal Statistical Society B*, 41, 190–195.
- Hansen, B. E. (1992). Testing for parameter instability in linear models. *Journal of Policy Modeling* 14, 517–533.
- Hendry, D. F. (1980). Econometrics: Alchemy or science? *Economica* 47, 387–406. Reprinted in Hendry, D. F. (1993), *Econometrics: Alchemy or Science?* Oxford: Blackwell Publishers.
- Hendry, D. F. (1995). *Dynamic Econometrics*. Oxford: Oxford University Press.
- Hendry, D. F. and J. A. Doornik (1996). *Empirical Econometric Modelling using PcGive for Windows*. London: Timberlake Consultants Press.
- Hendry, D. F. and H.-M. Krolzig (1999a). General-to-specific model specification using PcGets for Ox. Technical report, Institute of Economics and Statistics, Oxford University.
- Hendry, D. F. and H.-M. Krolzig (1999b). On the properties of general-to-simple modelling. Mimeo, Oxford Institute of Economics and Statistics, Oxford.

- Hoover, K. D. and S. J. Perez (1999). Data mining reconsidered: Encompassing and the general-to-specific approach to specification search. *Econometrics Journal* 2, 1–25.
- Judge, G. G. and M. E. Bock (1978). *The Statistical Implications of Pre-Test and Stein-Rule Estimators in Econometrics*. Amsterdam: North Holland Publishing Company.
- Krolzig, H.-M. and D. F. Hendry (1999). Computer automation of general-to-specific model selection procedures. Mimeo, Oxford Institute of Economics and Statistics, Oxford.
- Lovell, M. C. (1983). Data mining. *Review of Economics and Statistics* 65, 1–12.
- Mizon, G. E. (1977). Model selection procedures. In M. J. Artis and A. R. Nobay (Eds.), *Studies in Modern Economic Analysis*, pp. P97–120. Oxford: Basil Blackwell.
- Ramsey, J. B. (1969). Tests for specification errors in classical linear least squares regression analysis. *Journal of the Royal Statistical Society B* 31, 350–371.
- Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics* 6, 461–464.
- Sims, C. A., J. H. Stock, and M. W. Watson (1990). Inference in linear time series models with some unit roots. *Econometrica* 58, 113–144.
- White, H. (1980). A heteroskedastic-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica* 48, 817–838.
- Wooldridge, J. M. (1999). Asymptotic properties of some specification tests in linear models with integrated processes. In R. F. Engle and H. White (Eds.), *Cointegration, Causality and Forecasting*, pp. 366–384. Oxford: Oxford University Press.

10. APPENDIX

The variables are quarterly for the period 1960(3)–1988(3), where:

m_t	=	$\log M1$, seasonally adjusted.
p_t	=	$\log GNP$ deflator, seasonally adjusted (base 1982).
y_t	=	\log real GNP , seasonally adjusted.
$R_{l,t}$	=	20-year Treasury bond yield to maturity.
$R_{1,t}$	=	1-month Treasury bill coupon equivalent yield.
S_t	=	$R_{l,t} - R_{1,t}$.
AS_t^*	=	risk-adjusted average spread: $AS_t - 0.57V_t - 6.4\Delta SV_{t-1}$.
$R_{ma,t}$	=	learning-adjusted maximum yield on instruments in M2.
$R_{na,t}$	=	yield on NOW accounts.
$R_{sna,t}$	=	yield on super-NOW accounts.
$R_{nsa,t}$	=	learning-adjusted other-checkables rate in M1.
V_t	=	volatility measure based on long-bond holding-period yields.
SV_t	=	$\max(0, S_t) \cdot V_t$.
$\Delta_i x_t$	=	$(x_t - x_{t-i})$ for any variable x_t .
$\Delta^2 x_t$	=	$\Delta x_t - \Delta x_{t-1}$.
Ax_t	=	$(x_t + x_{t-1})$.
$\Delta \hat{p}_t$	=	$\Delta p_t + \Delta^2 p_t$.
$I1980_t$	=	Dummy for 1980 (2) = -1 ; 1980 (3) = $+1$.