New Developments in Automatic General-to-specific Modelling

David F. Hendry and Hans-Martin Krolzig*
Economics Department, Oxford University.

December 8, 2001

Abstract

We consider the analytic basis for *PcGets*, an Ox Package implementing automatic general-to-specific (*Gets*) modelling for linear, dynamic, regression models. *PcGets* mimics the theory of reduction – whereby empirical models arise – by commencing from a general congruent specification, which is then simplified to a minimal representation consistent with the desired selection criteria and the data evidence. We discuss the properties of *PcGets*, since results to date suggest that model selection can be relatively non-distortionary, even when the mechanism is unknown, and contrast Gets with possible alternatives.

1 Introduction

Empirical econometric modelling is an integral aspect of the attempt to make economics a quantitative science, although it raises many important methodological issues.[1] Not least among these is how to select models: by prior specification alone, by data evidence alone, or some mixture of these, where the 'weights' attached to the former could vary from unity to zero. Neither extreme seems tenable: economies are so high dimensional, non-stationary, complicated and subject to so many special events that pure theory could never precisely specify the underlying process; and there are simply too many variables to rely solely on data evidence. Thus, model-selection methods must be used, and the methodology thereof deserves careful attention.

When the prior specification of a possible relationship is not known for certain, data evidence is essential to delineate the relevant from the irrelevant variables. This is particularly true in non-stationary processes, as section 2 below notes, since retained irrelevant variables may be subject to deterministic shifts, inducing a structural break in the model under analysis. Thus, model selection is inevitable in practice; and while that may be accomplished in many possible ways, the chapter argues that simplification from a congruent general unrestricted model (GUM) embodies the best features of a number of alternatives. Some economists require the prior imposition of theory-based specifications: but such an approach assumes knowledge of the answer before the empirical investigation starts, so virtually denies empirical modelling any useful role – and in practice, it has rarely contributed in that framework (as argued by, e.g., Summers, 1991).

Unfortunately, there has been little agreement on which approaches should be adopted: see, *inter alia*, Leamer (1983b), Pagan (1987), Hendry, Leamer and Poirier (1990), Granger (1990) and Magnus and Morgan (1999). Hendry (2000b) discusses the rather pessimistic perceptions extant in the literature, including difficulties deriving from 'data-based model selection' (see the attack by Keynes, 1939, 1940,

*Prepared for *Econometrics and the Philosophy of Economics*, edited by Bernt P. Stigum. Financial support from the U.K. Economic and Social Research Council under grant L11625015 is gratefully acknowledged. We are indebted to Mike Clements, Jurgen Doornik, David Firth and especially Grayham Mizon for helpful comments.

[1]This chapter draws heavily on Hendry (2000b) and Hendry and Krolzig (2001).

on Tinbergen, 1940a, 1940b), 'measurement without theory' (Koopmans, 1947), 'data mining' (Lovell, 1983), pre-test biases' (Judge and Bock, 1978), 'ignoring selection effects' (Leamer, 1978), 'repeated testing' (Leamer, 1983a), 'arbitrary significance levels' (Hendry *et al.*, 1990), 'lack of identification' (see Faust and Whiteman, 1997, for a recent reiteration), and the potential 'path dependence of any selection' (Pagan, 1987). Nevertheless, none of these problems is inherently insurmountable, most are based on theoretical arguments (rather than evidence), and most have counter criticisms. Instead, the sequence of developments in automatic model selection initiated by Hoover and Perez (1999) suggests the converse: the operational characteristics of some selection algorithms are excellent across a wide range of states of nature, as this chapter will demonstrate. Naturally, we focus on *PcGets*: see Hendry and Krolzig (1999, 2001) and Krolzig and Hendry (2001).

PcGets is an Ox Package (see Doornik, 1999) implementing automatic general-to-specific (*Gets*) modelling for linear, dynamic, regression models based on the principles discussed in Hendry (1995). First, an initial general statistical model is tested for the absence of mis-specification (denoted congruence), and congruence is then maintained throughout the selection process by diagnostic checks, thereby ensuring a congruent final model. Next statistically-insignificant variables are eliminated by selection tests, both in blocks and individually. Many reduction paths are searched, to prevent the algorithm from becoming stuck in a sequence that inadvertently eliminates a variable which matters, and thereby retains other variables as proxies. Path searches in PcGets terminate when no variable meets the pre-set criteria, or any diagnostic test becomes significant. Non-rejected models are tested by encompassing: if several remain acceptable, so are congruent, undominated, mutually-encompassing representations, the reduction process recommences from their union, providing that is a reduction of the GUM, till a unique outcome is obtained: otherwise, or if all selected simplifications re-appear, the search is terminated using the Schwarz (1978) information criterion. Lastly, sub-sample insignificance seeks to identify 'spuriously significant' regressors.

In the Monte Carlo experiments in Hendry and Krolzig (1999) and Krolzig and Hendry (2001), PcGets recovers the data generation process (DGP) with an accuracy close to what one would expect if the DGP specification were known, but nevertheless coefficient tests were conducted. Empirically, on the data sets analyzed by Davidson, Hendry, Srba and Yeo (1978) and Hendry and Ericsson (1991b), PcGets selects (in seconds!) models at least as good as those developed over several years by their authors. Automatic model selection is in its infancy, yet already exceptional progress has been achieved, setting a high 'lower bound' on future performance. Moreover, there is a burgeoning symbiosis between the implementation and the theory – developments in each stimulate advances in the other.

The structure of the chapter is as follows. Section 2 considers the main alternative modelling approaches presently available, which strongly suggests a focus on *Gets* modelling. Since the theory of reduction is the basis for Gets modelling, section 3 notes the main ingredients of that theory. Section 4 then describes the embodiment of *Gets* in the computer program *PcGets* for automatic model selection. Section 5 discriminates between the costs of inference and the costs of search to provide an explanation of why PcGets does so well in model selection: sub-sections 5.2 and 5.3 then discuss the probabilities of deleting irrelevant, and selecting relevant, variables respectively. Section 6 presents a set of almost 2000 new experiments which have been used to calibrate the behaviour of the algorithm by varying the significance levels of its many selection procedures, then reports the outcomes of applying the released version of PcGets to both the Monte Carlo experiments in Lovell (1983) as re-analyzed by Hoover and Perez (1999), and in Krolzig and Hendry (2001), as well as noting the finite-sample behaviour of the mis-specification test statistics. Finally, section 7 concludes.

2 What are the alternatives?

Many investigators in econometrics have worried about the consequences of selecting models from data evidence, pre-dating even the Cowles Commission, as noted above. Eight literature strands can be delineated, which comprise distinct research strategies, if somewhat overlapping at the margins:

(1) simple-to-general modelling (see e.g., Anderson, 1962, Hendry and Mizon, 1978, and Hendry, 1979, for critiques);
(2) retaining the general model (see, e.g., Yancey and Judge, 1976, and Judge and Bock, 1978);
(3) testing theory-based models (see e.g., Hall, 1978, criticised by Davidson and Hendry, 1981, and Hendry and Mizon, 2000; Stigum, 1990 proposes a formal approach);
(4) other 'rules' for model selection, such as step-wise regression (see e.g., Leamer, 1983a, for a critical appraisal), and 'optimal' regression (see e.g., Coen, Gomme and Kendall, 1969, and the following discussion);
(5) model comparisons, often based on non-nested hypothesis tests or encompassing: see e.g., Cox (1961, 1962), Pesaran (1974), Deaton (1982), Kent (1986), and Vuong (1989), as well as Hendry and Richard (1982), Mizon (1984), Mizon and Richard (1986), and the survey in Hendry and Richard (1989);
(6) model selection by information criteria: see e.g., Schwarz (1978), Hannan and Quinn (1979), Amemiya (1980), Shibata (1980), Chow (1981), and Akaike (1985);
(7) Bayesian model comparisons: see e.g., Leamer (1978) and Clayton, Geisser and Jennings (1986);
(8) Gets: see e.g., Anderson (1962, 1971) for naturally ordered nested hypotheses, or more generally, Sargan (1973, 1981), Mizon (1977a, 1977b), Hendry (1979), and White (1990), with specific examples such as COMFAC (see Sargan, 1980), as well as the related literature on multiple hypothesis testing (well reviewed by Savin, 1984).

We briefly consider these in turn.

2.1 The problems of simple-to-general modelling

The paradigm of postulating a simple model and seeking to generalize it in the light of test rejections or anomalies is suspect for a number of reasons. First, there is no clear stopping point to an unordered search – the first non-rejection is obviously a bad strategy (see Anderson, 1962, 1971). Further, no control is offered over the significance level of testing, as it is not clear how many tests will be conducted.

Secondly, even if a general model is postulated at the outset as a definitive ending point, there remain difficulties with simple-to-general. Often, simple models are non-congruent, inducing multiple test rejections. When two or more statistics reject, it is unclear which (if either) has caused the problem; should both, or only one, be 'corrected'; or should other factors be sought? If several tests are computed *seriatim*, and a 'later' one rejects, then that invalidates all the earlier inferences, inducing a very inefficient research strategy. Indeed, until a model adequately characterizes the data, conventional tests are invalid: and it is obviously not sensible to skip testing in the hope that a model is valid.

Thirdly, alternative routes begin to multiply because simple-to-general is a divergent branching process – there are many possible generalizations, and the selected model evolves differently depending on which paths are selected, and in what order. Thus, genuine path dependence can be created by such a search strategy.

Fourthly, once a problem is revealed by a test, it is unclear how to proceed. It is a potentially dangerous *non sequitur* to adopt the alternative hypothesis of the test which rejected: e.g., assuming

that residual autocorrelation is error autoregression which imposes common factors on the dynamics (COMFAC: see Sargan, 1980, and Hendry and Mizon, 1978).

Finally, if a model displays symptoms of mis-specification, there is no point in imposing further restrictions on it. Thus, despite its obvious prevalence in empirical practice, there is little to commend a simple-to-general strategy.

2.2 Retaining the initial general model

Another alternative is to keep every variable in the GUM, but 'shrink' the parameter estimates (see, e.g., Yancey and Judge, 1976, and Judge and Bock, 1978). Shrinkage entails weighting coefficient estimates in relation to their significance, thus using a 'smooth' discount factor rather than the zero-one weights inherent in retain or delete decisions. Technically, the latter are 'inadmissible' (i.e., dominated on some criterion by other methods for a class of statistical problems). The shrinkage approach is also suggested as a solution to the 'pre-test' problem, whereby a biased estimator results when 'insignificant' draws are set to zero. Naturally, shrinkage also delivers biased estimators, but is argued to have a lower 'risk' than 'pre-test' estimators. However, such a claim has no theoretical underpinnings in processes subject to intermittent parameter shifts, since retaining irrelevant variables which are subject to deterministic shifts can be inimical to both forecasting and policy: see e.g., Clements and Hendry (2001). Moreover, progressivity entails explaining 'more by less', which such an approach hardly facilitates. Finally, absent omniscience, mis-specification testing is still essential to check the congruence of the GUM, which leads straight back to a testing – rather than a shrinkage – approach.

2.3 Testing theory models

Conventional econometric testing of economic theories is often conducted without first ascertaining the congruence of the models used: the dangers of doing so are discussed in Hendry and Mizon (2000). If, instead, mis-specification testing is undertaken, then decisions on how to react to rejections must be made. One response is to assume that the observed rejection can be nullified by using a suitably 'robust' inference procedure, such as heteroscedasticity-consistent standard errors as in White (1980), or also autocorrelation-consistent as in, say, Andrews (1991). Unfortunately, the symptoms of such problems may be due to other causes, such as unmodelled structural breaks, in which case the flaws in the inference are camouflaged rather than robustified. Another, earlier response was to generalize the initial model by assuming a model for the problem encountered (e.g., an autoregressive error when autocorrelated errors were found), which brings us back to the problems of simple-to-general. By protecting against such serious problems, PcGets may yet help 'data mining' to become a compliment.

2.4 Other model-simplification approaches

There are probably an almost uncountable number of ways in which models could be selected from data evidence (or otherwise). The only other 'rules' for model selection we consider here are step-wise regression (see e.g., Leamer, 1983a, for a critical appraisal), and 'optimal' regression (see e.g., Coen *et al.*, 1969, and the following discussion). There are three key problems with 'stepwise' regression. First, it does not check the congruence of the starting model, so cannot be sure the inference rules adopted have the operational characteristics believed of them (e.g., residual autocorrelation will distort estimated standard errors). Secondly, there are no checks on the congruence of reductions of the GUM, so again inference can become unreliable. Neither of these drawbacks is intrinsic to step-wise or 'optimal' regression (the latter tries almost every combination of variables, so borders on an information

criterion approach), and congruence testing could be added to those approaches. However, the key problem with step-wise is that only one simplification path is explored, usually based on successively eliminating the least-significant variable. Consequently, if a relevant variable is inadvertently eliminated early in the simplification, many others may be retained later to 'proxy' its role, so the algorithm can get stuck and select far too large a model: Hoover and Perez (1999) found that was an important problem in early versions of their algorithm for Gets, and hence implemented multiple-path searches.

Even in the ordered nested approach, a potential drawback is the existence of 'holes': e.g., when there exist seasonal autoregressive lags, intermediate parameters may be irrelevant, but will be retained if the ordering is imposed from the longest lag downwards. Thus, all these difficulties point to the importance of searching more feasible paths than those entertained by 'one route' methods.

2.5 Non-nested hypothesis tests and encompassing

Model comparisons can be based on non-nested hypothesis tests or encompassing: see e.g., Cox (1961, 1962), Pesaran (1974), Deaton (1982), and Vuong (1989) for the former, and Hendry and Richard (1982), Mizon (1984, 1995), Mizon and Richard (1986), Florens, Mouchart and Rolin (1990), Florens, Hendry and Richard (1996), Dhaene (1993), and the survey in Hendry and Richard (1989) for the latter. Links between these literatures are provided in Kent (1986), Wooldridge (1990), Govaerts, Hendry and Richard (1994), Hendry (1995), Marcellino (1996), and Lu and Mizon (2000). In such an approach, empirical models are developed by 'competing' investigators, and the 'winner' selected by non-nested, or encompassing, tests. Encompassing essentially requires a simple model to explain a more general one within which it is nested (often the union of the simple model with its rivals); this notion is called parsimonious encompassing and is denoted by $\underset{\surd}{\mathcal{E}}$. An important property of parsimonious encompassing is that it defines a partial ordering over models, since $\underset{\surd}{\mathcal{E}}$ is transitive, reflexive, and anti-symmetric. Since some aspects of inference must go wrong when a model is non-congruent, Gourieroux and Monfort (1995) argued for robust inference procedures. Alternatively, encompassing tests should only be conducted with respect to a 'baseline' congruent model, noting that non-congruence already demonstrates inadequacies: see Bontemps and Mizon (2001). Encompassing plays an important role in PcGets to select between congruent 'terminal' models found by different search paths.

2.6 Selecting models by minimizing an information criterion

Another route is to select models by minimizing an information criterion, especially a criterion which can be shown to lead to a consistent model selection (see e.g., Akaike, 1985, Schwarz, 1978, and Hannan and Quinn, 1979). These three 'selection' rules are denoted AIC (for the Akaike information criterion), SC (for the Schwarz criterion) and HQ (for Hannan–Quinn). The associated information criteria are defined as follows:

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

where L is the maximized likelihood, n is the number of estimated parameters and T is the sample size. The last two criteria ensure a consistent model selection: see e.g., Sawa (1978), Judge, Griffiths, Hill, Lütkepohl and Lee (1985), and Chow (1981). However, in 'unstructured problems', where there is no natural order to the hypotheses to be tested (see e.g., Mizon, 1977b), then a huge number of potential combinations must be investigated, namely 2^m possible models for m candidate variables: for the Lovell

(1983) database used below, that would be $2^{40} \simeq 10^{12}$ models, even restricting oneself to a maximum lag length of unity. To borrow an idea from Leamer (1983b), even at a 0.1 of a US cent per model, that would cost one billion US dollars. 'Shortcuts' such as dropping all but the k 'most significant' variables and exploring the remainder (as in Hansen, 1999), will work badly if more than k variables are needed. In practice, moreover, without checking that both the GUM and the selected model are congruent, the model which minimizes any information criterion has little to commend it (see e.g., Bontemps and Mizon, 2001). Information criteria also play an important role in PcGets to select when congruent and mutually-encompassing 'terminal' models are found from different search paths.

2.7 Bayesian model comparisons

Again, there are important overlaps in approach: for example, the Schwarz (1978) criterion is also known as the Bayesian information criterion (BIC). Perhaps the most enthusiastic proponent of a Bayesian approach to model selection is Leamer (1978, 1983b), leading to his 'practical' extreme-bounds analysis (see Leamer, 1983a, 1984, 1990), which in turn has been heavily criticized: see *inter alia*, McAleer, Pagan and Volker (1985), Breusch (1990) and Hendry and Mizon (1990). If there were a good empirical basis for non-distortionary 'prior' information, it would seem sensible to use it; but since one of the claims implicit in Leamer's critiques is that previous (non-Bayesian) empirical research is seriously flawed, it is difficult to see where such priors might originate. All too often, there is no sound basis for the prior, and a metaphysical element is introduced into the analysis.

2.8 Gets

General to simple approaches have a long pedigree: see *inter alia*, Anderson (1971), Sargan (1973, 1981), Mizon (1977b, 1977a), Hendry (1979), and White (1990). Important practical examples include COMFAC (see Sargan, 1980, and Hendry and Mizon, 1978), where the GUM must be sufficiently general to allow dynamic common factors to be ascertained; and cointegration (see e.g., Johansen, 1988), where the key feature is that all the inferences take place within the GUM (usually a vector autoregression). There is a very large related literature on testing multiple hypothesis (see e.g., Savin, 1984), where the consequences of alternative sequential tests are discussed.

2.9 Overview

PcGets blends aspects of all but the second-last of these many approaches, albeit usually with substantive modifications. For example, 'top down' searches (which retain the most significant variables, and block eliminate all others), partly resemble simple-to-general in the order of testing, but with the fundamental difference that the whole inference procedure is conducted within a congruent GUM. If the GUM cannot be reduced by the criteria set by the user, then it will be retained, though not weighted as suggested in the 'shrinkage' literature noted. The ability to fix some variables as always selected while evaluating the roles of others provides a viable alternative to simply including only the theoretically-relevant variables. Next, the key problems noted with 'stepwise' regression (that it only explores one path, so can get stuck, and does not check the congruence of either the starting model or reductions thereof) are avoided. Clearly, PcGets is a member of the Gets class, but also implements many pre-search reduction tests, and emphasizes block tests over individual when ever that is feasible. Although minimizing a model-selection criterion by itself does not check the congruence of the selection, which could therefore be rather poor, between mutually-encompassing congruent reductions, such criteria are applicable. Finally, parsimonious encompassing is used to select between congruent simplifications within a common GUM,

once contending terminal models have been chosen. Thus, we conclude that Gets is the most useful of the available approaches, and turn to a more formal analysis of its basis in the theory of reduction.

3 The theory of reduction

Gets is the practical embodiment of the theory of reduction: see e.g., Florens and Mouchart (1980), Hendry and Richard (1983), Florens *et al.* (1990), Hendry (1987), and Hendry (1995, ch. 9). That theory explains how the data generation process (DGP) characterizing an economy is reduced to the 'local' DGP (LDGP), which is the joint distribution of the subset of variables under analysis. By operations of aggregating, marginalizing, conditioning, sequentially factorizing, and transforming, a potentially vast initial information set is reduced to the small transformed subset that is relevant for the problem in question. These reduction operations affect the parameters of the process, and thereby determine the properties of the LDGP.

The theory of reduction not only explains the origins of the LDGP, the possible losses of information from any given reduction can be measured relative to its ability to deliver the parameters of interest in the analysis. For example, inappropriate reductions, such as marginalizing with respect to relevant variables, can induce non-constant parameters, or autocorrelated residuals, or heteroscedasticity and so on. The resulting taxonomy of possible information losses highlights six main null hypotheses against which model evaluation can be conducted, relating to the past, present, and future of the data, measurement and theory information, and results obtained by rival models. A *congruent* model is one that matches the data evidence on all the measured attributes: the DGP can always be written as a congruent representation. Congruence is checked in practice by computing a set of mis-specification statistics to ascertain that the residuals are approximately homoscedastic, normal, white noise, that any conditioning is valid, and that the parameters are constant. Empirical congruence is shown by satisfactory performance on all these checks. A model that can explain the findings of other models of the same dependent variables is said to encompass them, as discussed in section 2.5: see e.g., Cox (1961, 1962), Hendry and Richard (1982), Mizon (1984), Mizon and Richard (1986), and Hendry and Richard (1989). Since a model that explains the results of a more general model within which it is nested then parsimoniously encompasses the latter (see Govaerts *et al.*, 1994, and Florens *et al.*, 1996), when no reductions lose relevant information, the LDGP would parsimoniously encompass the DGP with respect to the subset of variables under analysis. Alternatively, a failure of the LDGP to parsimoniously encompass the DGP with respect to the parameters of interest suggests an inappropriate choice of variables for the analysis.

To implement *Gets*, a general unrestricted model (GUM) is formulated to provide a congruent approximation to the LDGP, given the theoretical and empirical background knowledge. Importantly, Bontemps and Mizon (2001) show that an empirical model is congruent if it parsimoniously encompasses the LDGP. The empirical analysis commences from this GUM, after testing for mis-specifications, and if no problems are apparent, the GUM is simplified to a parsimonious, congruent model, each simplification step being checked by diagnostic testing. Simplification can be done in many ways: and although the goodness of a model is intrinsic to it, and not a property of the selection route, poor routes seem unlikely to deliver useful models. Below, we investigate the impact of selection rules on the properties of the resulting models, and consider the solution proposed in Hoover and Perez (1999) of exploring many simplification paths.

4 The selection stages of PcGets

4.1 Formulating the GUM

Naturally, the specification of the initial general model must depend on the type of data (time series, cross section etc.), the size of sample, number of different potential variables, previous empirical and theoretical findings, likely functional-form transformations (e.g., logs) and appropriate parameterizations, known anomalies (such as measurement changes, breaks etc.) and data availability. The aim is to achieve a congruent starting point, so the specification should be sufficiently general that if a more general model is required, the investigator would be surprised, and therefore already have acquired useful information. Data may prove inadequate for the task, but even if a smaller GUM is enforced by pre-simplification, knowing at the outset what model ought to be postulated remains important. The larger the initial regressor set, the more likely adventitious effects are to be retained; but the smaller the GUM, the more likely key variables will be omitted. Further, the less orthogonality between variables, the more 'confusion' the algorithm faces, possibly leading to a proliferation of mutual-encompassing models, where final choices may only differ marginally (e.g., lag 2 versus 1).

Davidson and Hendry (1981, p.257) noted four possible problems facing Gets: (i) the chosen 'general' model may be inadequate, by being too special a case of the LDGP; (ii) data limitations may preclude specifying the desired relationship; (iii) the non-existence of an optimal sequence for simplification leaves open the choice of reduction path; and (iv) potentially-large type-II error probabilities of the individual tests may be needed to avoid a high type-I error of the overall sequence. By adopting the 'multiple path' development of Hoover and Perez (1999) (extended beyond the 10 paths they explored), and implementing a range of important improvements, PcGets overcomes the problems associated with points (iii) and (iv). However, the empirical success of PcGets depends crucially on the creativity of each researcher in specifying the general model, and the feasibility of estimating it from the available data – aspects beyond the capabilities of the program, other than the diagnostic tests serving their usual role of revealing model mis-specification.

There is a central role for economic theory in the modelling process in 'prior specification', 'prior simplification', and suggesting admissible data transforms. The first of these relates to the inclusion of potentially-relevant variables, the second to the exclusion of irrelevant effects, and the third to appropriate formulations in which the influences to be included are entered, such as log or ratio transforms etc., differences and cointegration vectors, and any likely linear transformations that might enhance orthogonality between regressors. The 'LSE approach' argued for a close link of theory and model, and explicitly opposed 'running regressions on every variable on the database' as in Lovell (1983) (see e.g., Hendry and Ericsson, 1991a). Unfortunately, economic theory rarely provides a basis for specifying the lag lengths in empirical macro-models: even when a theoretical model is dynamic, a 'time period' is usually not well defined. In practice, lags are chosen either for analytical convenience (e.g., first-order differential equation systems), or to allow for some desirable features (as in the choice of a linear, second-order difference equation to replicate cycles). Therefore, it seems sensible to start from an unrestricted autoregressive-distributed lag model with a maximal lag length set according to available evidence (e.g., as four or five lags for quarterly time series, to allow for seasonal dynamics). Prior analysis also remains essential for appropriate parameterizations; functional forms; choice of variables; lag lengths; and indicator variables (including seasonals, special events, etc.). Hopefully, automating the reduction process will enable researchers to concentrate their efforts on designing the GUM, which could significantly improve the empirical success of the algorithm.

4.1.1 Integrated variables

PcGets conducts all inferences as if the data are I(0). Most selection tests will in fact be valid even when the data are I(1), given the results in Sims, Stock and Watson (1990). Only t- or F-tests for an effect that corresponds to a unit root require non-standard critical values. The empirical examples on I(1) data noted in Krolzig and Hendry (2001) did not reveal problems, but in principle it could be useful to implement cointegration tests and appropriate transformations prior to reduction. Care is then required not to 'mix' variables with different degrees of integration, so our present recommendation is to specify the GUM in levels at least initially.

4.2 Mis-specification tests

Given the initial GUM, the next step is to conduct mis-specification tests. There must be sufficient tests to check the main attributes of congruence, but, as discussed above, not so many to induce a large type-I error. Thus, PcGets generally tests the following null hypotheses:

(1) white-noise errors;
(2) conditionally homoscedastic errors;
(3) normally distributed errors;
(4) unconditionally homoscedastic errors;
(5) constant parameters.

Approximate F-test formulations are used (see Harvey, 1981, and Kiviet, 1985, 1986). Section 6.4 describes the finite-sample behaviour of the various tests.

4.2.1 Significant mis-specification tests

If the initial mis-specification tests are significant at the pre-specified level, the required significance level is lowered, and search paths terminated only when that lower level is violated. Empirical investigators would probably re-specify the GUM on rejection, but as yet that relies on creativity beyond the capabilities of computer automation.

4.2.2 Integrated variables

Wooldridge (1999) shows that diagnostic tests on the GUM (and presumably simplifications thereof) remain valid even for integrated time series.

4.3 Pre-search reductions

Once congruence of the GUM is established, groups of variables are tested in the order of their absolute t-values, commencing from the smallest and continuing up towards the pre-assigned selection criterion, when deletion must become inadmissible. A non-stringent significance level is used at this step, usually 90%, since the insignificant variables are deleted permanently. Such a high value might seem surprising given the claim noted above that selection leads to over-parameterization, but confirms that such a claim is not sustainable. If no test is significant, the F-test on all variables in the GUM has been calculated, establishing that there is nothing to model.

Two rounds of cumulative simplification are offered, the second at a tighter level such as 25%. Optionally for time-series data, block tests of lag length are also offered.

4.4 Multiple search paths

All paths that commence with an insignificant t-deletion are explored. Blocks of variables constitute feasible search paths – like the block F-tests in the preceding sub-section, but along search paths – so these can be selected, in addition to individual-coefficient tests. Here we merely note that a non-null set of 'terminal' models is selected – namely all distinct minimal congruent reductions found along all the search paths – so when more than one such model is found, a choice between these is needed, accomplished as described in the next sub-section.

4.5 Encompassing

Encompassing tests select between the candidate congruent models at the end of 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, it is selected; otherwise, if some are rejected, PcGets forms the union of the remaining models, and repeats this round till no encompassing reductions result. That union then constitutes a new starting point, and the complete path-search algorithm repeats until the union is unchanged between successive rounds.

4.6 Information criteria

When such 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 preferred 'final-selection' rule presently is the Schwarz criterion, or BIC, defined above. For $T = 140$ and $m = 40$, minimum SC corresponds approximately to the marginal regressor satisfying $|\mathsf{t}| \geq 1.9$.

4.7 Sub-sample reliability

For the finally-selected model, sub-sample reliability is evaluated by the Hoover–Perez overlapping split-sample criterion. PcGets then concludes that some variables are definitely excluded; some definitely included; and some have an uncertain role, varying from a reliability of (say) 0% (included in the final model, but insignificantly, and insignificant in both sub-samples), through to 100% (significant overall and in both sub-samples). Investigators are at liberty to interpret such evidence as they see fit, noting that further simplification of the selected congruent model may induce some violations of congruence or encompassing.

Recursive estimation is central to the Gets research program, but focused on parameter constancy, whereas Hoover and Perez use the split samples to help determine overall significance. A central t-test wanders around the origin, so the probability is low that an effect which is significant only by chance in the full sample will also be significant in two independent sub-samples (see e.g., the discussion in Campos and Ericsson, 1999). Conversely, a non-central t-test diverges as the sample size increases, so should be significant in sub-samples, perhaps at a lower level of significance to reflect the smaller sample size. This strategy should be particularly powerful for model selection when breaks occur in some of the marginal relations over either of the sub-samples.

4.8 Type I and type II errors

Whether or not Gets over or under selects is not intrinsic to it, but depends on how it is used: neither type I nor type II errors are emphasized by the methodology *per se*, nor by the PcGets algorithm, but reflect the choices of critical values in the search process. In the Hendry and Krolzig (1999) analysis

of the Hoover and Perez (1999) re-run of the experiments in Lovell (1983), lowering the significance levels of the diagnostic tests from (say) 0.05 to 0.01 reduced the overall selection size noticeably (due to the difference in powering up 0.95 and 0.99 repeatedly), without greatly affecting the power of the model-selection procedure. Tighter significance levels (1% versus 5%) for diagnostic tests probably have much to commend them. Increasing the significance levels of the selection t-tests also reduced the empirical size, but lowered the power more noticeably for variables with population t-values smaller than 3. This trade-off can, therefore, be selected by an investigator. The next section addresses these issues.

5 Analyzing the algorithm

We first distinguish between the costs of inference and the costs of search, then consider some aspects of the search process.

5.1 Costs of inference and costs of search

Let $\mathsf{p}_{\alpha,i}^{dgp}$ denote the probability of retaining the i^{th} variable in the DGP when commencing from the DGP using a selection test procedure with significance level α. Then:

$$\sum_{i=1}^{k} \left(1 - \mathsf{p}_{\alpha,\mathsf{i}}^{\mathsf{dgp}}\right),$$

is a measure of the cost of inference when there are k variables in the DGP.

Let $\mathsf{p}_{\alpha,i}^{gum}$ denote the probability of retaining the i^{th} variable when commencing from the GUM, also using significance level α. Let $\mathcal{S}_{\nabla}$ denote the set of k relevant variables, and $\mathcal{S}_{\prime}$, the set of $(m-k)$ irrelevant, variables. Then pure search costs are:

$$\sum_{i \in \mathcal{S}_{\nabla}} \left(\mathsf{p}_{\alpha,\mathsf{i}}^{\mathsf{dgp}} - \mathsf{p}_{\alpha,\mathsf{i}}^{\mathsf{gum}}\right) + \sum_{i \in \mathcal{S}_{\prime}} \mathsf{p}_{\alpha,\mathsf{i}}^{\mathsf{gum}}.$$

For irrelevant variables, $\mathsf{p}_{\alpha,i}^{dgp} \equiv 0$, so the whole cost of retaining adventitiously-significant variables is attributed to search, plus any additional costs from failing to retain relevant variables. The former can be lowered by increasing the significance levels of selection tests, but at the cost of reducing the latter. However, it is feasible to lower size and raise power simultaneously by an improved search algorithm. When different selection strategies are used on the DGP and GUM (e.g., conventional t-testing on the former; pre-selection F-tests on the latter), then $\mathsf{p}_{\alpha,\mathsf{i}}^{\mathsf{gum}}$ could exceed $\mathsf{p}_{\alpha,\mathsf{i}}^{\mathsf{dgp}}$ (see, e.g., the critique of theory testing in Hendry and Mizon, 2000). We now consider the determinants of $\mathsf{p}_{\alpha,\mathsf{i}}^{\mathsf{gum}}$ for $i \in \mathcal{S}_{\prime}$, namely the non-deletion probabilities of irrelevant variables, then consider the probabilities $\mathsf{p}_{\alpha,\mathsf{i}}^{\mathsf{dgp}}$ of selecting the relevant variables assuming no irrelevant variables. The upshot of the analysis may surprise: the former can be made relatively small for reasonable critical values of the tests PcGets uses, and it is retaining the relevant variables that poses the real problem – or, the costs of search are small compared to the costs of inference.

5.2 Deletion probabilities

One might expect low deletion probabilities to entail high search costs when many variables are included but none actually matters. That would be true in a pure t-testing strategy as we now show: see Hendry

(2000a). The probabilities p_j of $j = 0, \ldots, m$ irrelevant variables being retained by t-testing are given by the j-th order coefficients in the expansion of $(\alpha + (1-\alpha))^m$, namely:

$$p_j = \frac{m!}{j!\,(m-j)!}\alpha^j\,(1-\alpha)^{m-j} \quad j = 0, \ldots, m. \tag{1}$$

Consequently, the expected number of variables retained in pure selection testing is:

$$r = \sum_{j=0}^{m} j p_j = m\alpha. \tag{2}$$

Per variable, this is a constant at $r/m = \alpha$. When $\alpha = 0.05$ and $m = 40$, r equals 2, falling to $r = 0.4$ for $\alpha = 0.01$: so even if only t-tests are used, few spurious variables would then be retained.

When there are no relevant variables, the probability of retaining none using only t-tests with critical value c_α is given by:

$$P\left(|t_i| < c_\alpha \;\forall i = 1, \ldots, m\right) = (1-\alpha)^m. \tag{3}$$

When $m = 40$ and $\alpha = 0.05$, there is only a 13% chance of retaining no spurious variables, but if $\alpha = 0.01$, this rises to a 67% chance of finding the null model. Even so, these constitute relatively low probabilities of locating the correct answer.

However, PcGets first calculates a one-off F-test, F_G, of the GUM against the null model using the critical value c_γ. Such a test has size $p_G = P(F_G \geq c_\gamma) = \gamma$ under the null, so the null model is selected with probability γ, which will be dramatically smaller than (3) even if γ is set at quite a high value, such as 0.10 (so the null is incorrectly rejected 10% of the time). Thus, other searches only commence $\gamma\%$ of the time, although (1) does not now describe the probabilities, as these are conditional on $F_G \geq c_\gamma$, so there is a higher probability of locating significant variables, even when $\alpha < \gamma$.

In the Hendry and Krolzig (1999) re-run of the Hoover–Perez experiments with $m = 40$, using $\gamma = 0.01$ yielded $\widehat{p_G} = 0.972$, as against a theory prediction from γ of 0.990. Let r variables be retained on t-testing using critical value c_α after the event $F_G \geq c_{0.01}$ occurs, then the probability $\overline{p}_r$ any given variable will be retained is $\overline{p}_r \simeq (1 - p_G)r/m$. The average 'non-deletion' probability across all the null-DGP Monte Carlos was $\widehat{\overline{p}}_r = 0.19\%$, so $r \simeq 3$ when $F_G \geq c_{0.01}$, and $r = 0$ otherwise. When a lower value of c_γ is set, using say $\gamma = 0.10$, the null model is retained 10% of the time, but providing the same c_α is used, fewer individually-significant variables will be retained in path searches, so the average size does not rise proportionally.

Overall, these are very small retention rates of spuriously-significant variables from a large number of irrelevant regressors, so it is easy to obtain a high probability of locating the null model even when 40 irrelevant variables are included, providing relatively tight significance levels are used – or a reasonably high probability using looser significance levels.

When only one, or a few, variables matter, the power of F_G becomes important, hence our suggesting somewhat less stringent critical values. As described above, when F_G rejects, PcGets next checks increasing simplifications of the GUM using the ordered values $t_1^2 \leq t_2^2 \leq \ldots \leq t_m^2$ in a cumulative F-test. Under orthogonality, we have the approximation after adding k variables:

$$F\left(k, T-k\right) \simeq \frac{1}{k}\sum_{i=1}^{k} t_i^2.$$

Once k variables are included, non-rejection requires that (a) $k-1$ variables did not induce rejection; (b) $|t_k| < c_\alpha$ for a critical value c_α; and (c) $F\left(k, T-k\right) < c_\gamma$ for a critical value c_γ. Slightly more than half the coefficients will have $t_i^2 < 0.5$ or smaller. Any $t_i^2 \leq 1$ reduces the mean F statistic, and

since $\mathsf{P}(|\mathsf{t}_i| < 1) = 0.68$, when $m = 40$ then approximately 28 variables fall in that group, leaving an F-statistic value of less than unity after their elimination. Also, $\mathsf{P}(|\mathsf{t}_i| \geq 2) = 0.05$ so only 2 out of 40 variables should chance to have a larger $|\mathsf{t}_i|$ value on average. Thus, surprisingly-large values of γ, such as 0.75, can be selected for this step yet have a high probability of eliminating most irrelevant variables. Since (e.g.) $\mathsf{P}(\mathsf{F}(30, 100) < 1|\mathsf{H}_0) \simeq 0.48$, a first step with $\gamma = 0.5$ would on average eliminate 28 variables with $\mathsf{t}_i^2 \leq 1$, when $m = 40$, and some larger t-values as well – hence the need to check that $|\mathsf{t}_k| < \mathsf{c}_\alpha$.

Further, a 'top down' search is also conducted, successively eliminating all but the largest t_m^2, and testing the remainder for zero, then all but the two largest etc. This works well when only a couple of variables matter.

Thus, in contrast to the high costs of inference to be demonstrated in the next section, the costs of search arising from retaining irrelevant variables seem relatively small. For a reasonable GUM, with say 20 variables where 15 are irrelevant, when using just t-tests at 5%, less than *one* spuriously-significant variable will be retained by chance. Pre-selection tests lower those probabilities. Against such costs, the next section shows that there is at most a 50% chance of retaining variables with non-centralities less than 2, and little chance of keeping several such regressors. Thus, the difficult problem is *retention of relevant, not elimination of irrelevant, variables*: critical values should be selected with these findings in mind. Practical usage of PcGets suggests that its operational characteristics are quite well described by this analysis: see section 6.

In applications, we often find that the multi-path searches and the pre-selection procedures produce similar outcomes, so although we cannot yet present a complete probability analysis of the former, it seems to behave almost as well in practice.

5.3 Selection probabilities

When searching a large database for a DGP, an investigator might retain the relevant regressors less often than when the correct specification is known, as well as retaining irrelevant variables in the finally-selected model. We now examine the difficulty of retaining relevant variables when commencing from the DGP, then turn to any additional power losses resulting from search.

Consider a t-test, denoted $\mathsf{t}(n, \psi)$, of a null hypothesis H_0 (where $\psi = 0$ under the null), when for a critical value c_α, a two-sided test is used with $\mathsf{P}(|\mathsf{t}(\mathsf{n}, 0)| \geq \mathsf{c}_\alpha|\mathsf{H}_0) = \alpha$. When the null is false, such a test will reject with a probability which varies with its non-centrality parameter ψ, c_α, and the degrees of freedom n. To calculate its power to reject the null when $\mathsf{E}[\mathsf{t}] = \psi > 0$, we approximate by:

$$\mathsf{P}(\mathsf{t} \geq \mathsf{c}_\alpha \mid \mathsf{E}[\mathsf{t}] = \psi) \simeq \mathsf{P}(\mathsf{t} - \psi \geq \mathsf{c}_\alpha - \psi \mid \mathsf{H}_0).$$

The following table from Hendry and Krolzig (2001) records some approximate power calculations when a single null hypothesis is tested and when six are tested, in each case, precisely once for $n = 100$ and different values of ψ and α, for a two-sided test.

t-test powers

ψ	α	$\mathsf{P}(\mathsf{t} \geq \mathsf{c}_\alpha)$	$\mathsf{P}(\mathsf{t} \geq \mathsf{c}_\alpha)^6$
1	0.05	0.165	0.000
2	0.05	0.508	0.017
2	0.01	0.267	0.000
3	0.05	0.845	0.364
3	0.01	0.646	0.073
4	0.01	0.914	0.583
5	0.01	0.990	0.941
6	0.001	0.995	0.968

The third column of the table reveals that there is little chance of retaining a variable with $\psi = 1$, and only a 50% chance of retaining a single variable with a population $|\mathsf{t}|$ of 2 when the critical value is also 2, falling to 25% for a critical value of 2.6. When $\psi = 3$, the power of detection is sharply higher for $\alpha = 0.05$, but still leads to more than 35% mis-classifications at $\alpha = 0.01$. Finally, when $\psi \geq 4$, one such variable will almost always be retained, even at stringent significance levels. Notice that no search is involved: these are the rejection probabilities of drawing from the correct univariate t-distribution, and merely reflect the vagaries of sampling. Further, the non-centrality must have a unique sign ($+$ or $-$), so only one side of the distribution matters under the alternative, although 2-sided nulls are assumed. These powers could be increased slightly by using a one-sided test when the sign is certain.

However, the final column shows that the probability of retaining all six relevant variables with the given non-centralities (i.e., the probability of locating the DGP) is essentially negligible when the tests are independent, except in the last three cases. Mixed cases (with different values of ψ) can be calculated by multiplying the probabilities in the third column (e.g., for $\psi = 2, 3, 3, 4, 5, 6$ the joint $\mathsf{P}(\cdot) \simeq 0.10$ at $\alpha = 0.01$). Such combined probabilities are highly non-linear in ψ, since one is almost certain to retain six variables with $\psi = 6$, even at a 0.1% significance level. The important conclusion is that, despite 'knowing' the DGP, low signal-noise variables will rarely be retained using t-tests when there is any need to test the null; and if there are many relevant variables, all of them are unlikely to be retained even when they have quite large non-centralities. The probabilities of retaining such variables when commencing from the GUM must be judged against this baseline, not against the requirement that the search procedure locate the 'truth'.

One alternative is to use an F-testing approach, after implementing (say) the first-stage pre-selection filters discussed above. A joint test will falsely reject a null model $\delta\%$ of the time when the implicit pre-selection critical value is c_δ, and the resulting model would then be the post-selection GUM. However, the reliability statistics should help reveal any problems with spuriously-significant variables. Conversely, this joint procedure has a dramatically higher probability of retaining a block of relevant variables. For example, if the 6 remaining variables all had expected t-values of two – an essentially impossible case above – then:

$$\mathsf{E}[\mathsf{F}(6, 100)] \simeq \frac{1}{6}\left(\sum_{\mathsf{i}=1}^{6} \mathsf{E}\left[\mathsf{t}_\mathsf{i}^2\right]\right) \simeq 4. \tag{4}$$

When $\delta = 0.025$, $c_\delta \simeq 2.5$ so to reject we need:

$$\mathsf{P}\left(\mathsf{F}(6, 100) \geq 2.5 \mid \overline{\mathsf{F}} = 4\right),$$

which we solve by using a non-central $\chi^2(6)$ approximation to $6\mathsf{F}(6,100)$ under the null, with critical value $c_{\alpha,k} = 14.5$, and the approximation under the alternative that:

$$\chi^2(6,24) = h\chi^2(m,0),$$

where:

$$h = \frac{6+48}{6+24} = 1.8 \text{ and } m = \frac{(6+24)^2}{6+48} \simeq 17, \tag{5}$$

so using:

$$\mathsf{P}\left[\mathsf{h}\chi^2(\mathsf{m},0) > \mathsf{c}_{\alpha,\mathsf{k}}\right] = \mathsf{P}\left[\chi^2(\mathsf{m},0) > \mathsf{h}^{-1}\mathsf{c}_{\alpha,\mathsf{k}}\right] \simeq \mathsf{P}\left[\chi^2(17,0) > 8\right] \simeq 0.97,$$

thereby almost always retaining all six relevant variables. This is in complete contrast with the near zero probability of retaining all six variables using t-tests on the DGP as above. Thus, the power-size trade-off depends on the search algorithm, and is not bounded above by the sequential t-test results.

The actual operating characteristics of PcGets are almost impossible to derive analytically given the complexity of the successive conditioning, so to investigate it in more detail, we consider some simulation studies.

6 Monte Carlo evidence on PcGets

Monte Carlo evidence on the behaviour of earlier incarnations of PcGets is presented in Hendry and Krolzig (1999) for the Hoover and Perez (1999) experiments based on Lovell (1983), and in Krolzig and Hendry (2001), together comprising two very different types of experimental design. Here we first consider a new set of almost 2000 experiments that we used to calibrate the behaviour of PcGets, then based on the resulting settings, we re-examine its performance on both earlier designs, noting the simulation behaviour of the mis-specification tests *en route*.

6.1 Calibrating PcGets by Monte Carlo experiments

We have also undertaken a huge range of Monte Carlo simulations across differing t-values in an artificial DGP, with different numbers of *relevant* regressors, different numbers of *irrelevant* regressors, and different critical values of most of the selection test procedures. The outcomes of these experiments were used to calibrate the in-built search strategies, which we denote by Liberal (minimize the non-selection probabilities) and Conservative (minimize the non-deletion probabilities).

The DGP was, for $m = k + n + 1$ variables:

$$\begin{aligned} y_t &= \sum_{i=1}^{k} \beta_{i,0} x_{i,t} + u_t, \qquad u_t \sim \mathsf{IN}[0,1], \\ \mathbf{x_t} &= \mathbf{v_t}, \qquad\qquad \mathbf{v_t} \sim \mathsf{IN}_{\mathsf{m+n}}[\mathbf{0}, \mathbf{I_{m+n}}]. \end{aligned} \tag{6}$$

As shown, only k of these generated variables entered the DGP with potentially non-zero coefficients. However, the GUM was:

$$y_t = \alpha y_{t-1} + \sum_{j=1}^{k+n} \beta_{j,i} x_{j,t} + \gamma + u_t. \tag{7}$$

We used a sample size of $T = 100$ and set the number of relevant variables, k to eight. There were only $M = 100$ replications in each experiment, but 1920 experiments in total. These comprised blocks

of 320 experiments across different settings of every selection-test significance level from a wide range, designed to highlight problem areas as well as good choices. In the first two blocks of experiments, all $\beta_{i,0} = 0$ so the null model was the DGP; in the next two, the expected t-value of each $\beta_{i,0}$ was 2, and in the final two blocks, the expected t-value of each $\beta_{i,0}$ was 3. The two blocks corresponded to generating n equal to eight or 20 additional irrelevant variables for inclusion in the GUM.

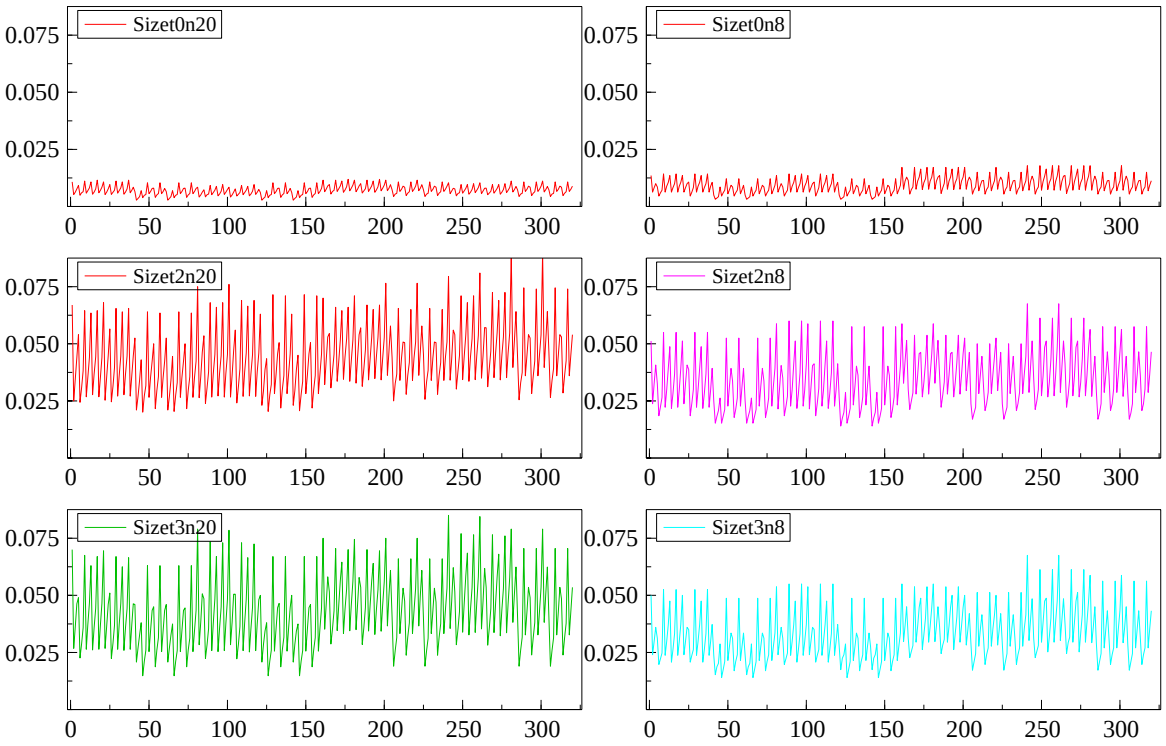


Figure 1 Null rejection frequencies.

The first figure has six plots, reporting the average null rejection frequency of each irrelevant variable included in the GUM, called the size for shorthand. As can be seen, the sizes in Figures 1a,b – when the DGP is the null model – are all quite small, around an average of 0.007–0.009 (the fluctuations in the graphs are probably due to sampling error from the small number of replications). Thus, the algorithm embodies 'automatic correction', in that although some significance levels are often set very loose, other aspects compensate to ensure a small overall size. For example, the F-test on the GUM against a null model has significance levels varying from 1% to 10%, with a negligible impact on the final sizes. More surprisingly, the average sizes per parameter in the GUM are higher for the smaller GUM. The increased null rejection frequency for $n = 8$ over $n = 20$ seems due to this property of the GUM and to pre-selection tests interacting with the number of variables retained when the null is rejected. Figures 1c–f show the corresponding sizes when there are relevant variables to be detected. The null rejection frequencies are higher in all states and for all critical value choices than for a null DGP, but range over 0.015–0.075 so remain relatively small.

Of course, the choice of critical value also affects power, so we now consider the power-size trade-offs for the four blocks of experiments with non-null DGPs. Figure 2 shows that for the experiments where $\mathsf{E}[\mathrm{t}] = 2$ for the 8 relevant variables, with both 20 and 8 irrelevant regressors. The latter always attains a higher power than the former, and comes close to the theoretical power of 0.5 applicable when drawing from a scalar t-distribution with a mean of 2 for $\alpha = 0.05$. Thus, the search induces only a small power loss at that size. The results for Fpgum at 10% and 5% illustrate the difficulties of deriving the characteristics of PcGets analytically: lowering its critical value from $c_{0.05}$ to $c_{0.10}$ leads to noticeably

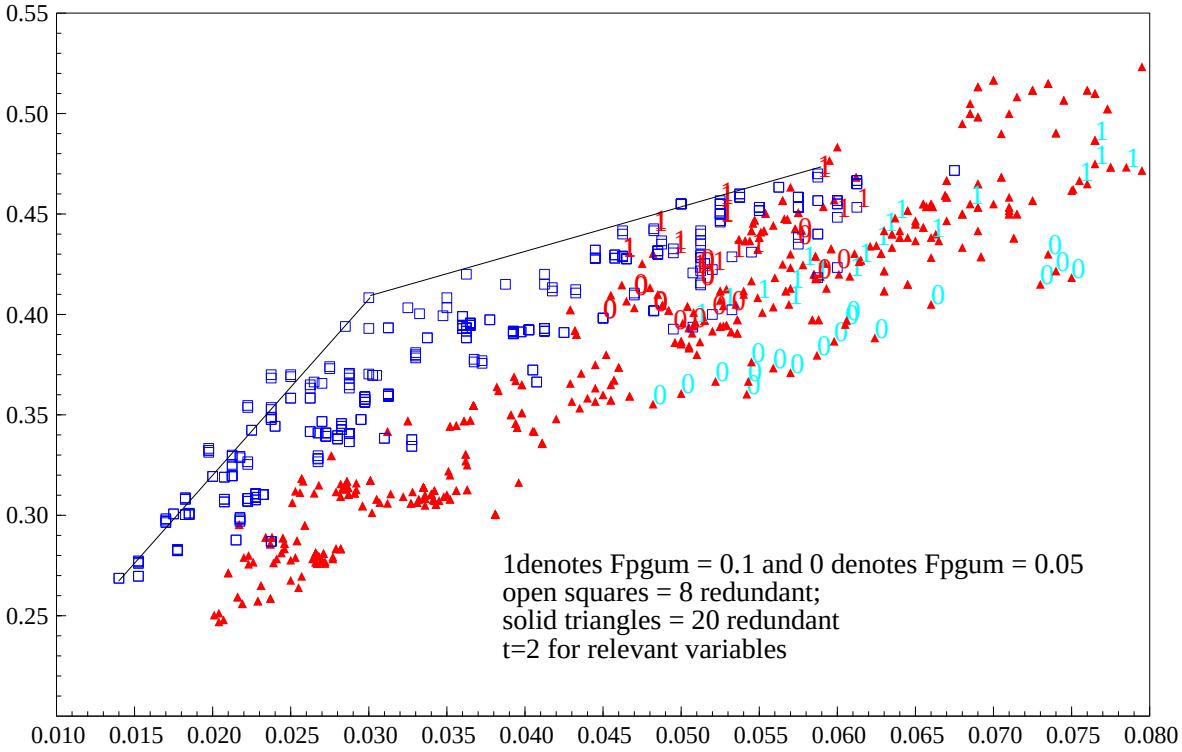


Figure 2 Power/size tradeoffs: t=2.

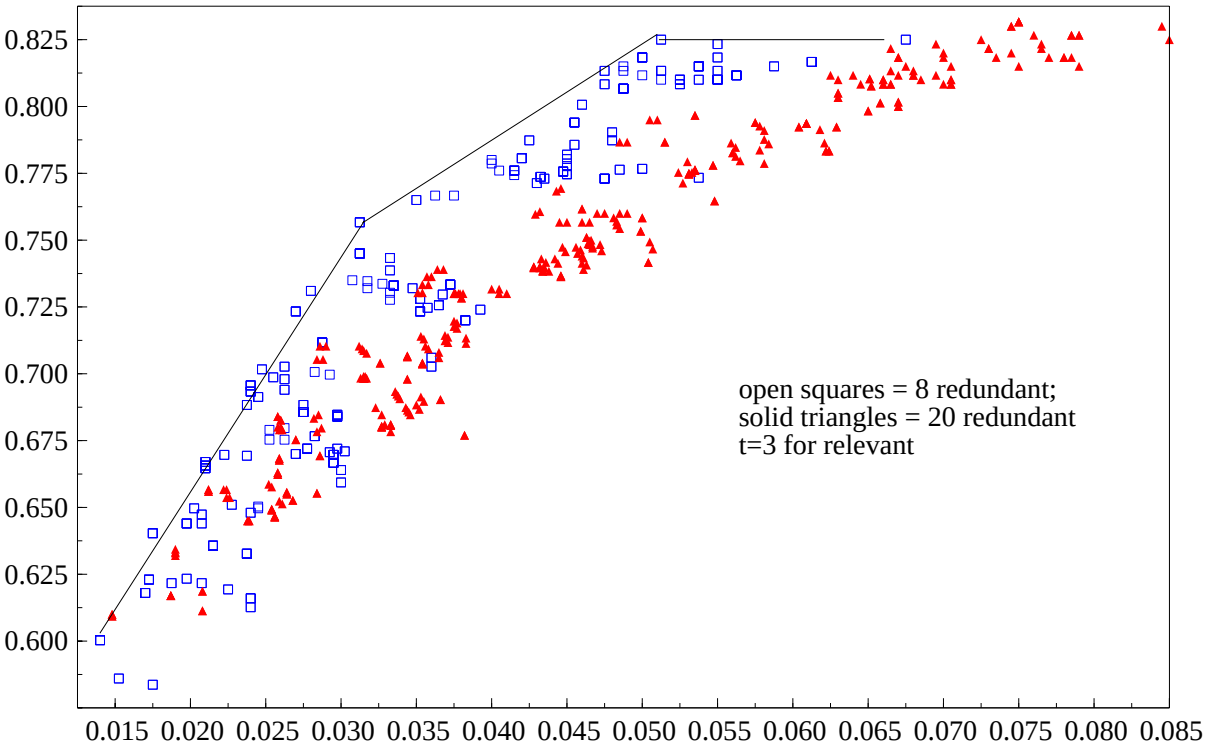


Figure 3 Power/size tradeoffs: t=3.

higher power with little increase in size. The ‘upper envelope’ (drawn by hand on the graphs) shows those combinations of significance levels that do best in this category of experiments for eight irrelevant variables, and guided our choice for the pre-set Liberal and Conservative strategies. Sizes smaller than 3% lead to relatively sharp falls in power, so are not recommended. Similarly, sizes much above 5%

yield small gains in power, so again seem dominated.

Similar considerations apply to the remaining experiments shown in Figure 3. Now the experiments with Fpgum at 10% and 5% are not shown separately as they were almost identical, supporting the selection of the former. Also, the theoretical power of 0.845 applicable when drawing from a scalar t-distribution with a mean of 3 for $\alpha = 0.05$ is nearly attained for 8 irrelevant variables, and is only slightly smaller for 20 irrelevant. The range from the worst to the best selection of critical values is about 0.05 for a given number of variables, so the algorithm is remarkably stable, more so for these larger t-values than in Figure 2.

Table 1 reports more detailed statistical outcomes for the settings we finally selected. 'T:' and 'S:' respectively denote the 'true' model (i.e., the DGP in (6)) and the specific model selected by PcGets in each replication. The evidence accords closely with the theory described above: when the population t-values are 2 or 3, it is almost impossible to retain the DGP specification even commencing from it when 1% critical values are used, and less than 25% likely for 5% critical values. Nevertheless, the actual sizes remain close to the nominal, and the powers for the specific model are close to the theoretical bound for the given choice of size. At 1%, the sizes (per variable) are lower for the larger GUM, probably due to the role of the block tests, which 'throw out' more variables when they fail to reject. We also record how the use of the reliability measure of significance in sub-samples to scale the outcomes would affect the size and power: i.e., if the reliability is 0.2 for a given variable, we treat an investigator as eliminating it 80% of the time. The reliability measure reduces both size and power, but the former more than the latter (as Hoover and Perez also found). Finally, the Conservative and Liberal strategies seem to have the correct operational characteristics.

6.2 Re-analyzing the 'Data Mining' experiments

Lovell (1983) formed a databank of 20 macro-economic variables; generated one (denoted y) as a function of zero to five others; regressed y on all others plus all lags thereof, four lags of y and an intercept; then examined how well some selection methods performed for the GUM:

$$y_t = \delta + \sum_{j=1}^{4} \alpha_j y_{t-j} + \sum_{i=1}^{18} \sum_{j=0}^{1} \gamma_{i,j} x_{i,t-j} + \omega_t. \tag{8}$$

He found none did even reasonably, but we suggest that was mainly because of flaws in the search algorithms, not the principle of selection.

In retrospect, despite using actual macroeconomic data, the Lovell experiments are not very representative of real situations likely to confront econometricians, for four reasons. First, the few variables which matter the most have (absolute) t-values of 5, 8, 10 and 12 in the population, so are almost always jointly detected, irrespective of the significance level set: even using $\alpha = 0.1\%$ only requires $|\mathsf{t}| > 3.4$. Secondly, the remaining relevant variables have population t-values of less than unity, so will almost never be detected:

$$\mathsf{P}\left(|\mathsf{t}| \geq 2 \mid \mathsf{E}\left[\mathsf{t}\right] = 1\right) \simeq \mathsf{P}\left(\mathsf{t} \geq 2 \mid \mathsf{E}\left[\mathsf{t}\right] = 1\right) = \mathsf{P}\left(\mathsf{t} \geq 1 \mid \mathsf{E}\left[\mathsf{t}\right] = 0\right),$$

which is less than 16% even at $\alpha = 5\%$, and about 5% at $\alpha = 1\%$. Thus, there is essentially a zero probability of retaining two such variables in those experiments (and hence no chance of locating the DGP), *even when no search is involved*. Thirdly, including more than **40** irrelevant variables when the sample size is $T = 100$ is hardly representative of empirical modelling. Finally, and true of most such Monte Carlo experiments, the DGP is a special case of the GUM, so mis-specification tests play no useful role.

Table 1 Liberal and Conservative strategies.

DGP Design	A	B	C	D	E	F
t	0	0	2	2	3	3
k	8	8	8	8	8	8
n	8	20	8	20	8	20
conservative						
T:DGPfound	1.0000	1.0000	0.0000	0.0000	0.0230	0.0280
S:DGPfound	0.8430	0.7870	0.0000	0.0000	0.0250	0.0230
S:NonDeletion	0.1570	0.2130	0.1780	0.2630	0.1800	0.2870
S:NonSelection	0.0000	0.0000	1.0000	1.0000	0.9690	0.9720
T:Dominated (0.025)	0.1560	0.2120	0.4730	0.4280	0.3530	0.3060
S:Dominated (0.025)	0.0010	0.0010	0.4270	0.4300	0.5160	0.4890
S:Size	0.0108	0.0093	0.0223	0.0166	0.0229	0.0178
S:Power	——	——	0.2800	0.2575	0.6091	0.5820
S:Selection error	——	——	0.3712	0.2240	0.2069	0.1321
reliability based						
S:Size	0.0090	0.0084	0.0144	0.0131	0.0147	0.0141
S:Power	——	——	0.2398	0.2355	0.5539	0.5489
S:Selection error	——	——	0.3873	0.2278	0.2304	0.1390
liberal						
T:DGPfound	1.0000	1.0000	0.0230	0.0050	0.1740	0.2000
S:DGPfound	0.4330	0.3590	0.0250	0.0010	0.1130	0.0680
S:NonDeletion	0.5670	0.6410	0.1800	0.6920	0.4570	0.6700
S:NonSelection	0.0000	0.0000	0.9690	0.9970	0.8220	0.8380
T:Dominated (0.075)	0.5670	0.6410	0.3530	0.7550	0.6270	0.6050
S:Dominated (0.075)	0.0000	0.0000	0.5160	0.0760	0.1470	0.1070
S:Size	0.0524	0.0479	0.0229	0.0627	0.0650	0.0595
S:Power	——	——	0.6091	0.4638	0.7900	0.7702
S:Selection error	——	——	0.2069	0.1980	0.1375	0.1082
reliability based						
S:Size	0.0464	0.0429	0.0147	0.0556	0.0553	0.0530
S:Power	——	——	0.5539	0.4373	0.7559	0.7436
S:Selection error	——	——	0.2304	0.2005	0.1497	0.1111

Combining these facets, any researcher running, or re-running, such experiments knows this aspect, so is 'biased' towards setting tough selection rules, and ignoring diagnostic checks: see e.g., the approach in Hansen (1999), commenting on Hoover and Perez (1999) – PcGets would do best with very stringent significance levels. Unfortunately, in many practical settings, such settings will not perform well: t-values of 2 or 3 will rarely be retained, and badly mis-specified models would not be detected. Consequently, we kept the critical value settings suggested by, and all the diagnostic tests used in, the calibration experiments described in section 6.1.

Table 2 records the DGPs in those experiments that did not involve variables with population t-statistics less than unity in absolute value. In all cases, $\varepsilon_t \sim \mathsf{IN}[0,1]$. The GUM nested the DGP, with the addition of between 37–40 irrelevant variables, depending on the experiment.

Table 2 Selected Hoover–Perez DGPs.

HP1	$y_t = 130\varepsilon_t$
HP2	$y_t = 0.75y_{t-1} + 130\varepsilon_t$
HP2*	$y_t = 0.50y_{t-1} + 130\varepsilon_t$
HP7	$y_t = 0.75y_{t-1} + 1.33x_t - 0.975x_{t-1} + 9.73\varepsilon_t$

Note: the dependent variable choice differs across experiments.

Table 3 DGP t-values in Hoover–Perez experiments.

Experiment	HP1	HP2	HP2*	HP7
y_{t-1}	–	12.95	4.70	12.49
x_t	–	–	–	15.14
x_{t-1}	–	–	–	-8.16

First, we record for convenience the original outcomes reported for our re-run of the Hoover–Perez experiments, shown in table 4 (HP3 was not one of those we re-ran). While the performance was sometimes spectacular – as in HP1, where the DGP (which is the null model) is almost always found – it could also be less satisfactory, as in HP7 with $\alpha = 0.05$.

Table 4 Original outcomes for Hoover–Perez experiments.

Experiment	HP1	HP2	HP2	HP7	HP7
Significance level	0.01	0.01	0.01*	0.01	0.05
Selection probabilities					
y_{t-1}		1.0000	1.0000	1.0000	1.0000
x_t				1.0000	1.0000
x_{t-1}				0.9970	0.9980
Power	——	1.0000	1.0000	0.9990	0.9990
Size	0.0019	0.0242	0.0088	0.0243	0.1017
Selected Model					
DGP found	0.9720	0.6020	0.8520	0.5900	0.1050
Non-DGP var. included	0.0280	0.3980	0.1480	0.4100	0.8950
DGP var. not included	0.0000	0.0000	0.0000	0.0030	0.0020
DGP is dominated	0.0260	0.3830	0.1030	0.3900	0.8900
Specific is dominated	0.0020	0.0150	0.0450	0.0200	0.0050

The basic PcGets setting used in table 5 was the Liberal strategy, with some variation in the two

Table 5 Simulation results for Hoover–Perez experiments.

PcGets	Significance levels		
F-test of GUM	0.2500	0.5000	0.5000
Pre-search (stage I)	0.5000	0.2500	0.5000
HP1	probability	probability	probability
T:DGPfound	1.0000	1.0000	1.0000
S:DGPfound	0.8980	0.8800	0.8530
S:NonDeletion	0.1020	0.1200	0.1470
S:NonSelection	——	——	——
T:Dominated	0.1020	0.1200	0.1470
S:Dominated	0.0000	0.0000	0.0000
S:Size	0.0049	0.0049	0.0064
S:Power	——	——	——
S:Size (rel)	0.0041	0.0040	0.0054
S:Power (rel)	——	——	——
HP2	probability	probability	probability
T:DGPfound	1.0000	1.0000	1.0000
S:DGPfound	0.8090	0.8570	0.8100
S:NonDeletion	0.1910	0.1430	0.1900
S:NonSelection	0.0010	0.0000	0.0000
T:Dominated	0.1860	0.1430	0.1860
S:Dominated	0.0040	0.0000	0.0040
S:Size	0.0095	0.0073	0.0095
S:Power	0.9990	1.0000	1.0000
S:Size (rel)	0.0079	0.0058	0.0079
S:Power (rel)	0.9990	1.0000	1.0000
HP2*	probability	probability	probability
T:DGPfound	0.9940	0.9940	0.9940
S:DGPfound	0.6990	0.7710	0.7590
S:NonDeletion	0.2140	0.1780	0.2080
S:NonSelection	0.1100	0.0860	0.0490
T:Dominated	0.1960	0.1470	0.1970
S:Dominated	0.0850	0.0490	0.0310
S:Size	0.0104	0.0084	0.0102
S:Power	0.8900	0.9140	0.9510
S:Size (rel)	0.0088	0.0070	0.0086
S:Power (rel)	0.8894	0.9127	0.9501
HP7	probability	probability	probability
T:DGPfound	1.0000	1.0000	1.0000
S:DGPfound	0.8080	0.8460	0.8080
S:NonDeletion	0.1920	0.1540	0.1920
S:NonSelection	0.0060	0.0070	0.0060
T:Dominated	0.1850	0.1440	0.1850
S:Dominated	0.0010	0.0030	0.0010
S:Size	0.0096	0.0074	0.0096
S:Power	0.9980	0.9977	0.9980
S:Size (rel)	0.0081	0.0062	0.0081
S:Power (rel)	0.9979	0.9976	0.9979

pre-selection tests, namely the F-test on the GUM being the null model (denoted Fpgum), and the F-test for the significance of the lag length (Fplag). The outcomes are based on $M = 1000$ replications of the DGP with a sample size of $T = 100$.

The probabilities of retaining the DGP when commencing from it, and from the GUM (denoted T:DGPfound and S:DGPfound) are shown first: the former is always close to unity and the latter almost always above 75% for the range of experiments shown. The power of PcGets (the probability of retaining the variables that matter) is close to that when commencing from the DGP, and the size is usually less than 0.75%, even using the Liberal strategy – although with more than 37 irrelevant variables, the Conservative strategy would almost certainly be a better choice.

Next the non-deletion and non-selection probabilities are shown: the latter is usually tiny, so the former is close to 1−S:DGPfound. Finally, T:Dominated and S:Dominated record the probabilities that the DGP or the selected model dominates (i.e., encompasses) the other: as can be seen, the former occurs quite often, between 13–20%, whereas the latter is usually under 5%. Thus, the operating characteristics are stable between the experiments, and everywhere well behaved.

Table 6 Re-running the Hoover–Perez experiments.

Experiment	HP1	HP2	HP2*	HP7
conservative				
T:DGPfound	1.0000	1.0000	0.9940	1.0000
S:DGPfound	0.8530	0.8100	0.7590	0.8080
S:NonDeletion	0.1470	0.1900	0.2080	0.1920
S:NonSelection	——	0.0000	0.0490	0.0060
T:Dominated (0.025)	0.1470	0.1860	0.1970	0.1850
S:Dominated (0.025)	0.0000	0.0040	0.0310	0.0010
S:Size	0.0064	0.0095	0.0102	0.0096
S:Power	——	1.0000	0.9510	0.9980
reliability based				
S:Size	0.0054	0.0079	0.0086	0.0081
S:Power	——	1.0000	0.9501	0.9979
liberal				
T:DGPfound	1.0000	1.0000	0.9990	1.0000
S:DGPfound	0.4700	0.3450	0.3440	0.3470
S:NonDeletion	0.5300	0.6550	0.6500	0.6530
S:NonSelection	——	0.0000	0.0220	0.0070
T:Dominated (0.075)	0.5290	0.6540	0.6380	0.6440
S:Dominated (0.075)	0.0010	0.0010	0.0050	0.0020
S:Size	0.0427	0.0549	0.0551	0.0541
S:Power	——	1.0000	0.9780	0.9977
reliability based				
S:Size	0.0377	0.0476	0.0482	0.0470
S:Power	——	1.0000	0.9762	0.9976

Finally, we show how the latest version of PcGets (May, 2001) would perform using Conservative settings: see table 6. The probabilities of locating the DGP are now constant across states of nature, and remain high: this is due to both high power to retain relevant, and a small size for irrelevant, variables. The reliability measure is denoted (rel): again the size is clearly reduced, being everywhere less than 1%, without much loss of power, confirming the value of the reliability check.

Overall, these finding cohere with those reported earlier (for a different version of the program,

and very different settings for the significance levels), and suggest that PcGets performs well even in a demanding problem, where the GUM is hugely over-parameterized. The outcomes suggest that relatively loose critical values should be chosen for pre-selection tests.

6.3 Rerunning the JEDC experiments

In this final set of experiments from Krolzig and Hendry (2001), the DGP is a Gaussian regression model, where the strongly-exogenous variables are Gaussian white-noise processes:

$$\begin{aligned} y_t &= \sum_{k=1}^{5} \beta_{k,0} x_{k,t} + \varepsilon_t, \qquad \varepsilon_t \sim \mathsf{IN}\left[0,1\right], \\ x_t &= v_t, \qquad\qquad\qquad\qquad v_t \sim \mathsf{IN}_{10}\left[\mathbf{0}, \mathbf{I_{10}}\right] \quad \text{for } \mathsf{t} = 1,\ldots,\mathsf{T}, \end{aligned} \tag{9}$$

where $\beta_{1,0} = 2/\sqrt{T}$, $\beta_{2,0} = 3/\sqrt{T}$, $\beta_{3,0} = 4/\sqrt{T}$, $\beta_{4,0} = 6/\sqrt{T}$, $\beta_{5,0} = 8/\sqrt{T}$.

The GUM is an $ADL(1,1)$ model which includes as non-DGP variables the lagged endogenous variable y_{t-1}, the strongly-exogenous variables $x_{6,t},\ldots,x_{10,t}$ and the first lags of all regressors:

$$y_t = \pi_{0,1} y_{t-1} + \sum_{k=1}^{10}\sum_{i=0}^{1} \pi_{k,i} x_{k,t-i} + \pi_{0,0} + u_t, \;\; u_t \sim \mathsf{IN}\left[0,\sigma^2\right]. \tag{10}$$

The sample size used here is just $T = 100$, and the number of replications M is 1000: the non-zero population t-values are therefore $2,3,4,6,8$. In (10), 17 of 22 regressors are 'nuisance'.

We record the performance of the original PcGets and that on the calibrated settings now embodied in the two formal strategies. The progress is obvious: the sizes are generally similar with higher powers, again close to the upper bound of drawing from a scalar t-distribution. The search costs are generally negligible when compared to the costs of statistical inference. Under the default setting of PcGets (i.e., pre-search and split-sample analysis are active), the costs of search are reduced to 0.0015 and 0.0054 per variable at nominal sizes of 0.01 and 0.05. Thus, they are just 2.12% (respectively 8.49%) of the underlying costs of statistical inference.

Ensuring empirical congruence by checking the validity of the reductions is almost free: in the case of the liberal strategy, the diagnostic tests contributed only an estimated 0.0004 to the total costs of search which are 0.0015. Partly, this is because the mis-specification tests are well behaved, as we now discuss.

6.4 Small-sample properties of the mis-specification tests

Krolzig and Hendry (2001) recorded quantile–quantile (QQ) plots of the empirical distributions of seven potential mis-specification tests when estimating the correct specification, the general model, and the finally-selected model in a Monte Carlo study with $T = 1000$ and $T = 100$. In the larger sample, the tests were unaffected by the presence of (strongly-exogenous) nuisance regressors (i.e., variables that entered the GUM but not the DGP). However, at the smaller sample size, some strong deviations were evident from the theoretical distributions for two tests: the portmanteau statistic (see Box and Pierce, 1970) rejected serial independence of the errors too often in the correct specification, never in the general, and too rarely in the final model, and the hetero-x test (see White, 1980) often had degrees-of-freedom problems for the GUM, but anyway performed badly for both the true and final models. Such incorrect finite-sample sizes of these two diagnostic tests would induce excess rejections of congruent GUMs, resulting in an increased overall size: thus, the portmanteau and the hetero-x diagnostics should be excluded from any battery of test statistics. Overall, therefore, we recommend the five mis-specification tests shown in table 9.

Table 7 JEDC Experiments: Conservative Strategy.

	Theory	JEDC	JEDC	PcGets	PcGets	PcGets
presearch		——	yes	——	yes	yes
spit sample		——	——	——	——	yes
Size	0.0100	0.0185	0.0088	0.0162	0.0135	0.0102
vs. JEDC			-0.0097	-0.0023	-0.0050	-0.0084
vs. JEDC (%)			*-52.43*	*-12.43*	*-27.19*	*-45.14*
Power	0.7544	0.7598	0.6665	0.7582	0.7490	0.7312
vs. JEDC			-0.0933	-0.0016	-0.0108	-0.0286
vs. JEDC (%)			*-12.28*	*-0.21*	*-1.42*	*-3.77*
Selection error	0.0635	0.0689	0.0826	0.0675	0.0675	0.0689
vs. JEDC			0.0137	-0.0014	-0.0014	0.0001
vs. JEDC (%)			*19.89*	*-2.05*	*-2.08*	*0.08*
Costs of search		0.0053	0.0190	0.0039	0.0039	0.0054
Costs of search (%)		*8.40*	*29.97*	*6.18*	*6.15*	*8.49*
Expected number of						
NonDGP variables incl.	0.17	0.31	0.15	0.28	0.23	0.17
DGP variables deleted	1.23	1.20	1.67	1.21	1.26	1.34
variables misplaced	1.40	1.52	1.82	1.48	1.48	1.52
Power function						
power (t=2)	0.2580	0.2820	0.1538	0.2850	0.2820	0.2559
power (t=3)	0.6130	0.6200	0.4278	0.6240	0.6190	0.5925
power (t=4)	0.9020	0.8980	0.7645	0.8820	0.8720	0.8577
power (t=6)	0.9990	0.9990	0.9865	1.0000	1.0000	0.9994
power (t=8)	1.0000	1.0000	1.0000	1.0000	1.0000	0.9997
Indicators						
T:DGPfound				0.1540	0.1540	0.1540
S:DGPfound				0.1290	0.1290	0.1290
S:NonDeletion				0.2130	0.1890	0.1890
S:NonSelection				0.8320	0.8390	0.8390
T:Dominated				0.6630	0.6490	0.6490
S:Dominated				0.1680	0.1850	0.1850

Table 8 JEDC Experiments: Liberal Strategy.

	Theory	JEDC	JEDC	PcGets	PcGets	PcGets
presearch		——	yes	——	yes	yes
spit sample		——	——	——	——	yes
Size	0.0500	0.0677	0.0477	0.0666	0.0546	0.0447
vs. JEDC			-0.0200	-0.0011	-0.0131	-0.0230
vs. JEDC (%)			*-29.54*	*-1.62*	*-19.36*	*-33.99*
Power	0.8522	0.8532	0.8156	0.8556	0.8446	0.8274
vs. JEDC			-0.0376	0.0024	-0.0086	-0.0258
vs. JEDC (%)			*-4.41*	*0.28*	*-1.01*	*-3.02*
Selection error	0.0722	0.0857	0.0788	0.0843	0.0775	0.0738
vs. JEDC			-0.0069	-0.0014	-0.0082	-0.0119
vs. JEDC (%)			*-8.06*	*-1.63*	*-9.54*	*-13.91*
Costs of search		0.0135	0.0065	0.0121	0.0053	0.0015
Costs of search (%)		*18.62*	*9.06*	*16.69*	*7.30*	*2.12*
Expected number of						
NonDGP variables incl.	0.85	1.15	0.81	1.13	0.93	0.76
DGP variables deleted	0.74	0.73	0.92	0.72	0.78	0.86
variables misplaced	1.59	1.88	1.73	1.85	1.71	1.62
Power function						
power (t=2)	0.4730	0.4930	0.4080	0.5000	0.4820	0.4494
power (t=3)	0.8120	0.8020	0.7330	0.8180	0.8080	0.7798
power (t=4)	0.9760	0.9720	0.9390	0.9600	0.9580	0.9455
power (t=6)	1.0000	0.9990	0.9980	1.0000	1.0000	1.0000
power (t=8)	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
Indicators						
T:DGPfound				0.3960	0.3960	0.3960
S:DGPfound				0.1410	0.1700	0.1700
S:NonDeletion				0.6290	0.5780	0.5780
S:NonSelection				0.5960	0.6150	0.6150
T:Dominated				0.7550	0.7000	0.7000
S:Dominated				0.0470	0.0520	0.0520

Table 9 Test battery.

Test	*Alternative*	*Statistic*	*Sources*
Chow(τT)	Predictive failure over a subset of $(1-\tau)T$ obs.	$\mathsf{F}((1-\tau)\mathsf{T}, \tau\mathsf{T}-\mathsf{k})$	Chow (1960, p.594-595), Hendry (1979)
normality test	Skewness and excess kurtosis	$\chi^2(2)$	Jarque and Bera (1980), Doornik and Hansen (1994)
AR 1-p test	$p-$th order residual autocorrelation	$\mathsf{F}(\mathsf{p}, \mathsf{T}-\mathsf{k}-\mathsf{p})$	Godfrey (1978), Harvey (1981, p.173)
ARCH 1-p test	$p-$th order autoregressive conditional heteroskedasticity	$\mathsf{F}(\mathsf{p}, \mathsf{T}-\mathsf{k}-\mathsf{p})$	Engle (1982), Engle, Hendry and Trumbull (1985)
hetero test	Heteroscedasticity quadratic in regressors x_i^2	$\mathsf{F}(\mathsf{q}, \mathsf{T}-\mathsf{k}-\mathsf{q}-1)$	White (1980), Nicholls and Pagan (1983)

There are T observations and k regressors in the model under the null. The value of q may differ across statistics, as may those of k and T across models. By default, *PcGets* sets $p = 4$, $r = 12$, and computes two Chow tests at $\tau_1 = [0.5T]/T$ and $\tau_2 = [0.9T]/T$.

We investigated the empirical distributions of these five remaining test statistics in the DGP, GUM and selected model for the experiments in section 6.3. The results are shown in Figure 4, and demonstrate that the test distributions in the selected model are both satisfactory and close to the distributions of those tests under the null of the DGP. The properties of the first three mis-specification tests were also satisfactory in the GUM, but the ARCH and hetero tests then departed notably from their theoretical distributions. Thus, care should be exercised when applying these to heavily over-parametrized models. One possible strategy would be apply such tests after the pre-search reductions had eliminated 'highly irrelevant' variables.

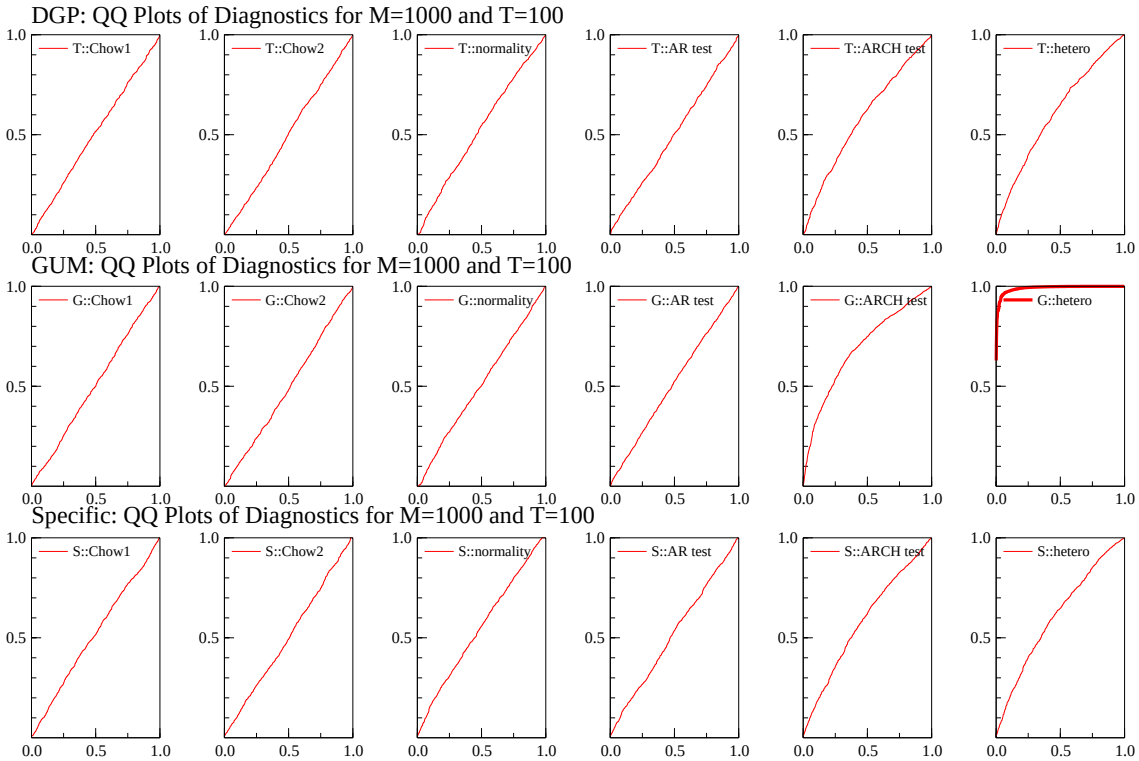


Figure 4 Selecting mis-specification tests: QQ plots for $T = 100$.

7 Conclusion

In his survey of model selection for the *Handbook of Econometrics*, Ed Leamer deems the need for model selection to be a serious drawback for economics research:

> "*On pessimistic days I doubt that economists have learned anything from the mountains of computer print-outs that fill their offices. On especially pessimistic days, I doubt that they ever will.*"

A lot has changed since he wrote in the early 1980s, not least the demise of 'mountains of computer print-outs'. We have demonstrated that automatic model selection procedures can perform almost as well as could be hoped, with size close to the desired level and power near that applicable when commencing from the DGP. Thus, the problem is not one of search costs due to 'the absence of completely defined models' (*pace* Leamer), but to the properties of statistical inference.

However, here too we are more optimistic – at least in time-series applications, new data continually accrue, helping to weed out adventitiously-significant variables, and confirming the relevance of the remainder. Model selection is part of a progressive research strategy, not a one-off attempt to forge 'economic laws' that hold for ever.

Certainly, the context of regression with strongly exogenous variables is far too simple to characterize real-world econometrics. Empirical researchers confront non-stationary, mis-measured data, on evolving dynamic and high-dimensional economies, with at best weakly exogenous conditioning variables. At the practical level, Gets is applicable to systems, such as vector autoregressions (see Krolzig, 2000), and endogenous regressors where sufficient instruments exist. And it is just as powerful a tool on cross-section problems, as demonstrated by Hoover and Perez (2000).

Moreover, the study of automatic selection procedures has barely begun. As Hendry and Mizon (2000) remark, "*early chess-playing programs were easily defeated, but later ones can systematically beat Grandmasters, so we anticipate computer-automated model selection software developing well beyond the capabilities of the most expert modellers.*" Deep Blue may be just round the corner.

References

Akaike, H. (1985). Prediction and entropy. In Atkinson, A. C., and Fienberg, S. E. (eds.), *A Celebration of Statistics*, pp. 1–24. New York: Springer-Verlag.

Amemiya, T. (1980). Selection of regressors. *International Economic Review*, **21**, 331–354.

Anderson, T. W. (1962). The choice of the degree of a polynomial regression as a multiple-decision problem. *Annals of Mathematical Statistics*, **33**, 255–265.

Anderson, T. W. (1971). *The Statistical Analysis of Time Series.* New York: John Wiley & Sons.

Andrews, D. W. K. (1991). Heteroskedasticity and autocorrelation consistent covariance matrix estimation. *Econometrica*, **59**, 817–858.

Bontemps, C., and Mizon, G. E. (2001). Congruence and encompassing. In Stigum, B. P. (ed.), *Econometrics and the Philosophy of Economics*. Princeton: Princeton University Press. forthcoming.

Box, G. E. P., and Pierce, D. A. (1970). Distribution of residual autocorrelations in autoregressive-integrated moving average time series models. *Journal of the American Statistical Association*, **65**, 1509–1526.

Breusch, T. S. (1990). Simplified extreme bounds. in Granger (1990), pp. 72–81.

Campos, J., and Ericsson, N. R. (1999). Constructive data mining: Modeling consumers' expenditure in Venezuela. *Econometrics Journal*, **2**, 226–240.

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

Chow, G. C. (1981). Selection of econometric models by the information criteria. In Charatsis, E. G. (ed.), *Proceedings of the Econometric Society European Meeting 1979*, Ch. 8. Amsterdam: North-Holland.

Clayton, M. K., Geisser, S., and Jennings, D. E. (1986). A comparison of several model selection procedures. In Goel, P., and Zellner, A. (eds.), *Bayesian Inference and Decision Techniques*: Elsevier Science.

Clements, M. P., and Hendry, D. F. (2001). Modelling methodology and forecast failure. *Econometrics Journal*, Forthcoming.

Coen, P. G., Gomme, E. D., and Kendall, M. G. (1969). Lagged relationships in economic forecasting. *Journal of the Royal Statistical Society A*, **132**, 133–163.

Cox, D. R. (1961). Tests of separate families of hypotheses. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 1, pp. 105–123 Berkeley: University of California Press.

Cox, D. R. (1962). Further results on tests of separate families of hypotheses. *Journal of the Royal Statistical Society B*, **24**, 406–424.

Davidson, J. E. H., and Hendry, D. F. (1981). Interpreting econometric evidence: The behaviour of consumers' expenditure in the UK. *European Economic Review*, **16**, 177–192. Reprinted in Hendry, D. F., (1993), (2000a) *op. cit.*

Davidson, J. E. H., Hendry, D. F., Srba, F., and Yeo, J. S. (1978). Econometric modelling of the aggregate time-series relationship between consumers' expenditure and income in the United Kingdom. *Economic Journal*, **88**, 661–692. Reprinted in Hendry, D. F., (1993), (2000a) *op. cit.*

Deaton, A. S. (1982). Model selection procedures or, does the consumption function exist. In Chow, G. C., and Corsi, P. (eds.), *Evaluating the Reliability of Macro-Economic Models*, Ch. 5. New York: John Wiley.

Dhaene, G. (1993). Encompassing: Formulation, properties, and testing. Unpublished thesis, Leuven, Belgium.

Doornik, J. A. (1999). *Object-Oriented Matrix Programming using Ox*. London: Timberlake Consultants Press. 3rd edition.

Doornik, J. A., and Hansen, H. (1994). A practical test for univariate and multivariate normality. Discussion paper, Nuffield College.

Engle, R. F. (1982). Autoregressive conditional heteroskedasticity with estimates of the variance of UK inflation. *Econometrica*, **50**, 987–1008.

Engle, R. F., Hendry, D. F., and Trumbull, D. (1985). Small sample properties of ARCH estimators and tests. *Canadian Journal of Economics*, **43**, 66–93.

Faust, J., and Whiteman, C. H. (1997). General-to-specific procedures for fitting a data-admissible, theory-inspired, congruent, parsimonious, encompassing, weakly-exogenous, identified, structural model of the DGP: A translation and critique. *Carnegie–Rochester Conference Series on Public Policy*, **47**, 121–161.

Florens, J.-P., Hendry, D. F., and Richard, J.-F. (1996). Encompassing and specificity. *Econometric*

Theory, **12**, 620–656.

Florens, J.-P., and Mouchart, M. (1980). Initial and sequential reduction of Bayesian experiments. Discussion paper 8015, CORE, Louvain-La-Neuve, Belgium.

Florens, J.-P., Mouchart, M., and Rolin, J.-M. (1990). *Elements of Bayesian Statistics.* New York: Marcel Dekker.

Godfrey, L. G. (1978). Testing for higher order serial correlation in regression equations when the regressors include lagged dependent variables. *Econometrica*, **46**, 1303–1313.

Gourieroux, C., and Monfort, A. (1995). Testing, encompassing, and simulating dynamic econometric models. *Econometric Theory*, **11**, 195–228.

Govaerts, B., Hendry, D. F., and Richard, J.-F. (1994). Encompassing in stationary linear dynamic models. *Journal of Econometrics*, **63**, 245–270.

Granger, C. W. J. (ed.)(1990). *Modelling Economic Series*. Oxford: Clarendon Press.

Hall, R. E. (1978). Stochastic implications of the life cycle-permanent income hypothesis: Evidence. *Journal of Political Economy*, **86**, 971–987.

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

Hansen, B. E. (1999). Discussion of '*data mining reconsidered*'. *Econometrics Journal*, **2**, 26–40.

Harvey, A. C. (1981). *The Econometric Analysis of Time Series*. Deddington: Philip Allan.

Hendry, D. F. (1979). Predictive failure and econometric modelling in macro-economics: The transactions demand for money. In Ormerod, P. (ed.), *Economic Modelling*, pp. 217–242. London: Heinemann. Reprinted in Hendry, D. F., (1993), (2000a) *op. cit.*

Hendry, D. F. (1987). Econometric methodology: A personal perspective. In Bewley, T. F. (ed.), *Advances in Econometrics*, Ch. 10. Cambridge: Cambridge University Press.

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

Hendry, D. F. (2000a). *Econometrics: Alchemy or Science?* Oxford: Oxford University Press. New Edition.

Hendry, D. F. (2000b). Epilogue: The success of general-to-specific model selection. In *Econometrics: Alchemy or Science?*, pp. 467–490. Oxford: Oxford University Press. New Edition.

Hendry, D. F., and Ericsson, N. R. (1991a). An econometric analysis of UK money demand in 'Monetary Trends in the United States and the United Kingdom' by Milton Friedman and Anna J. Schwartz. *American Economic Review*, **81**, 8–38.

Hendry, D. F., and Ericsson, N. R. (1991b). Modeling the demand for narrow money in the United Kingdom and the United States. *European Economic Review*, **35**, 833–886.

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

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

Hendry, D. F., Leamer, E. E., and Poirier, D. J. (1990). A conversation on econometric methodology. *Econometric Theory*, **6**, 171–261.

Hendry, D. F., and Mizon, G. E. (1978). Serial correlation as a convenient simplification, not a nuisance: A comment on a study of the demand for money by the Bank of England. *Economic Journal*, **88**, 549–563. Reprinted in Hendry, D. F., (1993), (2000a) *op. cit.*

Hendry, D. F., and Mizon, G. E. (1990). Procrustean econometrics: or stretching and squeezing data. in Granger (1990), pp. 121–136.

Hendry, D. F., and Mizon, G. E. (2000). Reformulating empirical macro-econometric modelling. *Oxford Review of Economic Policy*, **16**, 138–159.

Hendry, D. F., and Richard, J.-F. (1982). On the formulation of empirical models in dynamic econometrics. *Journal of Econometrics*, **20**, 3–33. Reprinted in Granger, C. W. J. (ed.) (1990), *op. cit.*, Oxford: Clarendon Press and in Hendry, D. F., (1993), (2000a) *op. cit.*

Hendry, D. F., and Richard, J.-F. (1983). The econometric analysis of economic time series (with discussion). *International Statistical Review*, **51**, 111–163. Reprinted in Hendry, D. F., (1993), (2000a) *op. cit.*

Hendry, D. F., and Richard, J.-F. (1989). Recent developments in the theory of encompassing. In Cornet, B., and Tulkens, H. (eds.), *Contributions to Operations Research and Economics. The XXth Anniversary of CORE*, pp. 393–440. Cambridge, MA: MIT Press.

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

Hoover, K. D., and Perez, S. J. (2000). Truth and robustness in cross-country growth regressions. unpublished paper, Economics Department, University of California, Davis.

Jarque, C. M., and Bera, A. K. (1980). Efficient tests for normality, homoscedasticity and serial independence of regression residuals. *Economics Letters*, **6**, 255–259.

Johansen, S. (1988). Statistical analysis of cointegration vectors. *Journal of Economic Dynamics and Control*, **12**, 231–254. Reprinted in R.F. Engle and C.W.J. Granger (eds), Long-Run Economic Relationships, Oxford: Oxford University Press, 1991, 131–52.

Judge, G. G., and Bock, M. E. (1978). *The Statistical Implications of Pre-Test and Stein-Rule Estimators in Econometrics*. Amsterdam: North Holland Publishing Company.

Judge, G. G., Griffiths, W. E., Hill, R. C., Lütkepohl, H., and Lee, T.-C. (1985). *The Theory and Practice of Econometrics*, 2nd edn. New York: John Wiley.

Kent, J. T. (1986). The underlying nature of nonnested hypothesis tests. *Biometrika*, **73**, 333–343.

Keynes, J. M. (1939). Professor Tinbergen's method. *Economic Journal*, **44**, 558–568.

Keynes, J. M. (1940). Comment. *Economic Journal*, **50**, 154–156.

Kiviet, J. F. (1985). Model selection test procedures in a single linear equation of a dynamic simultaneous system and their defects in small samples. *Journal of Econometrics*, **28**, 327–362.

Kiviet, J. F. (1986). On the rigor of some mis-specification tests for modelling dynamic relationships. *Review of Economic Studies*, **53**, 241–261.

Koopmans, T. C. (1947). Measurement without theory. *Review of Economics and Statistics*, **29**, 161–179.

Krolzig, H.-M. (2000). General-to-specific reductions in vector autoregressive processes. Economics discussion paper, 2000-w34, Nuffield College, Oxford.

Krolzig, H.-M., and Hendry, D. F. (2001). Computer automation of general-to-specific model selection procedures. *Journal of Economic Dynamics and Control*, **25**, 831–866.

Leamer, E. E. (1978). *Specification Searches. Ad-Hoc Inference with Non-Experimental Data*. New York: John Wiley.

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

Reprinted in Granger, C. W. J. (ed.) (1990), *op. cit.*

Leamer, E. E. (1983b). Model choice and specification analysis. In Griliches, Z., and Intriligator, M. D. (eds.), *Handbook of Econometrics*, Vol. 1, Ch. 5. Amsterdam: North-Holland.

Leamer, E. E. (1984). Global sensitivity results for generalized least squares estimates. *Journal of the American Statistical Association*, **79**, 867–870.

Leamer, E. E. (1990). Sensitivity analyses would help. in Granger (1990), pp. 88–96.

Lovell, M. C. (1983). Data mining. *Review of Economics and Statistics*, **65**, 1–12.

Lu, M., and Mizon, G. E. (2000). Nested models, orthogonal projection, and encompassing. In Marriott, P., and Salmon, M. (eds.), *Applications of Differential Geometry to Econometrics*, pp. 64–84. Cambridge: Cambridge University Press. ISBN 0 521 651166.

Magnus, J. R., and Morgan, M. S. (eds.)(1999). *Methodology and Tacit Knowledge: Two Experiments in Econometrics*. Chichester: John Wiley and Sons.

Marcellino, M. (1996). Encompassing and exogeneity. Unpublished phd thesis, European University Institute.

McAleer, M., Pagan, A. R., and Volker, P. A. (1985). What will take the con out of econometrics?. *American Economic Review*, **95**, 293–301. Reprinted in Granger, C. W. J. (ed.) (1990), *op. cit.*

Mizon, G. E. (1977a). Inferential procedures in nonlinear models: An application in a UK industrial cross section study of factor substitution and returns to scale. *Econometrica*, **45**, 1221–1242.

Mizon, G. E. (1977b). Model selection procedures. In Artis, M. J., and Nobay, A. R. (eds.), *Studies in Modern Economic Analysis*, pp. 97–120. Oxford: Basil Blackwell.

Mizon, G. E. (1984). The encompassing approach in econometrics. In Hendry, D. F., and Wallis, K. F. (eds.), *Econometrics and Quantitative Economics*, pp. 135–172. Oxford: Basil Blackwell.

Mizon, G. E. (1995). Progressive modelling of macroeconomic time series: the LSE methodology. In Hoover, K. D. (ed.), *Macroeconometrics: Developments, Tensions and Prospects*, pp. 107–169. Dordrecht: Kluwer Academic Press.

Mizon, G. E., and Richard, J.-F. (1986). The encompassing principle and its application to non-nested hypothesis tests. *Econometrica*, **54**, 657–678.

Nicholls, D. F., and Pagan, A. R. (1983). Heteroscedasticity in models with lagged dependent variables. *Econometrica*, **51**, 1233–1242.

Pagan, A. R. (1987). Three econometric methodologies: A critical appraisal. *Journal of Economic Surveys*, **1**, 3–24. Reprinted in Granger, C. W. J. (ed.) (1990), *op. cit.*

Pesaran, M. H. (1974). On the general problem of model selection. *Review of Economic Studies*, **41**, 153–171.

Sargan, J. D. (1973). Model building and data mining. Presented to the Association of University Teachers of Economics, Meeting, Manchester. *Econometric Reviews*, (2001), **20**, 159–170.

Sargan, J. D. (1980). Some tests of dynamic specification for a single equation. *Econometrica*, **48**, 879–897. Reprinted as pp. 191–212 in Sargan J. D. (1988), *Contributions to Econometrics*, Vol. 1, Cambridge: Cambridge University Press.

Sargan, J. D. (1981). The choice between sets of regressors. Mimeo, Economics Department, London School of Economics. *Econometric Reviews*, (2001), **20**, 171–186.

Savin, N. E. (1984). Multiple hypothesis testing. In Griliches, Z., and Intriligator, M. D. (eds.), *Handbook of Econometrics*, Vol. 2–3, Ch. 14. Amsterdam: North-Holland.

Sawa, T. (1978). Information criteria for discriminating among alternative regression models. *Econometrica*, **46**, 1273–1292.

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

Shibata, R. (1980). Asymptotically efficient selection of the order of the model for estimating parameters of a linear process. *Annals of Statistics*, **8**, 147–164.

Sims, C. A., Stock, J. H., and Watson, M. W. (1990). Inference in linear time series models with some unit roots. *Econometrica*, **58**, 113–144.

Stigum, B. P. (1990). *Towards a Formal Science of Economics*. Cambridge, MA: MIT Press.

Summers, L. H. (1991). The scientific illusion in empirical macroeconomics. *Scandinavian Journal of Economics*, **93**, 129–148.

Tinbergen, J. (1940a). *Statistical Testing of Business-Cycle Theories*. Geneva: League of Nations. Vol. I: A Method and its application to Investment Activity.

Tinbergen, J. (1940b). *Statistical Testing of Business-Cycle Theories*. Geneva: League of Nations. Vol. II: Business Cycles in the United States of America, 1919–1932.

Vuong, Q. H. (1989). Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica*, **57**, 307–333.

White, H. (1980). A heteroskedastic-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, **48**, 817–838.

White, H. (1990). A consistent model selection. in Granger (1990), pp. 369–383.

Wooldridge, J. M. (1990). An encompassing approach to conditional mean tests with applications to testing nonnested hypotheses. *Journal of Econometrics*, **45**, 331–350.

Wooldridge, J. M. (1999). Asymptotic properties of some specification tests in linear models with integrated processes. In Engle, R. F., and White, H. (eds.), *Cointegration, Causality and Forecasting*, pp. 366–384. Oxford: Oxford University Press.

Yancey, T. A., and Judge, G. G. (1976). A Monte Carlo comparison of traditional and stein-rule estimators under squared error loss. *Journal of Econometrics*, **4**, 285–294.