The Properties of Automatic *Gets* Modelling

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

First version: September 2002
This version: March 2003

Abstract

We examine the properties of automatic model selection, as embodied in *PcGets*, and evaluate its performance across different (unknown) states of nature. After describing the basic algorithm and some recent changes, we discuss the consistency of its selection procedures, then examine the extent to which model selection is non-distortionary at relevant sample sizes. The problems posed in judging performance on collinear data are noted. The conclusion notes how *PcGets* can handle more variables than observations, and hence how it can tackle non-linear models.

1 Introduction

Model selection theory poses great difficulties: all statistics for selecting models and evaluating their specifications have distributions, usually interdependent, different under null and alternative, and altered by every modelling decision. Fortunately, recent advances in computer automation of selection algorithms have allowed a fresh look at this old problem, by allowing operational studies of alternative strategies: see *inter alia* Hoover and Perez (1999) and Krolzig and Hendry (2001). An overview of the literature, and the developments leading to general-to-specific (*Gets*) modelling in particular, is provided by Campos, Ericsson and Hendry (2003). Here we analyze the properties of *PcGets*, and seek to ascertain its behaviour in sifting relevant from irrelevant variables in econometric modelling.[1] Hendry and Krolzig (2003b) describe the selection strategies embodied in *PcGets*, their foundation in the theory of reduction, and potential alternatives. They emphasize the distinction between the costs of inference, which are an inevitable consequence of non-zero significance levels and non-unit powers, versus the costs of search, which are additional to those faced when commencing from a model that is the data generation process (DGP). Finally, they calibrate its settings by Monte Carlo.

The structure of this paper is as follows. After outlining its algorithm, we describe some recent changes in section 2. Then section 3 considers the large-sample performance of *PcGets* through the consistency of its selection procedures, as embodied in its pre-programmed Liberal and Conservative strategies: section 3.1 compares it with model selection based on information criteria (see e.g., Schwarz, 1978). Thirdly, its finite-sample behaviour is examined in section 4 across a range of Monte Carlo experiments from Hendry and Krolzig (1999, 2003b) and Krolzig and Hendry (2001). Next, section 5 investigates possible small-sample 'pre-test biases' and 'model-selection effects' for both estimators and tests in a single experiment. Section 6 comments on the impact of collinearity on selection probabilities. Section 7 concludes, noting two new developments: the first when there are more variables than

*We are indebted to Dorian Owen and Julia Campos for helpful comments and suggestions.

[1] *PcGets* by Hendry and Krolzig (2001) is an Ox Package (see Doornik, 2001) implementing automatic general-to-specific (*Gets*) modelling for linear regression models based on the theory of reduction, as in Hendry (1995, Ch.9).

observations, which surprisingly, is not necessarily a major problem for *PcGets*; and secondly, selecting a non-linear model when the desired class is known.

2 The selection algorithm

PcGets has four basic stages in its approach to selecting a parsimonious undominated representation of an overly general initial model, denoted the general unrestricted model (GUM). The first concerns the estimation and testing of the GUM (1–4 below), the second is the pre-search process (5–6); the third is the multi-path search procedure (7–13); and the fourth is the post-search evaluation (14). The following sketches the main steps involved: see Hendry and Krolzig (2001) for details.

(1) formulate the GUM, based on theory, previous evidence, and institutional knowledge, seeking a relatively orthogonal parameterization;
(2) select the set of mis-specification tests to be checked and their forms (e.g., residual autocorrelation of r^{th}-order);
(3) set the significance levels of all selection and mis-specification tests to ensure the desired overall null rejection frequency, perhaps by selecting one of the pre-set strategies;
(4) check that the GUM captures the essential characteristics of the data (congruence), perhaps with outlier adjustments;
(5) undertake pre-search reduction tests at a loose significance level (these include lag-order pre-selection, F-tests on successively shorter lag groups and cumulative F-tests based on t-tests ordered from the smallest up, and the largest down);
(6) eliminate insignificant variables to reduce the search complexity, and estimate the new GUM as the baseline for the next stage;
(7) multiple reduction path searches now commence from each feasible initial deletion (to avoid path-dependent selections);
(8) diagnostically check the validity of each reduction, to ensure the congruence of the final model;
(9) if all reduction and diagnostic tests are acceptable and all remaining variables are significant (or further reductions induce mis-specifications), that model becomes a *terminal* selection, and the next path search commences;
(10) when all paths have been explored and all distinct terminal models have been found, they are repeatedly tested against their union to find an encompassing contender;
(11) rejected models are removed, and the union of the 'surviving' terminal models becomes the GUM of a repeated multi-path search iteration;
(12) this entire search process continues till a unique choice emerges, or search converges to a set of mutually encompassing and undominated contenders;
(13) in that last case, all the selected models are reported and a unique final choice made by the desired selection criterion;
(14) the significance of every variable in the *final* model is assessed in two over-lapping sub-samples to check the reliability of the selection.

Several changes to this basic algorithm have been implemented since Hendry and Krolzig (2001), so we briefly describe these. Most only slightly altered the program's behaviour, reflecting how near the theoretical upper bound performance already is, and the degree of 'error correction' manifest in the experiments used to calibrate the program (when one procedure performed poorly, another usually did well). Nevertheless, improvements remain feasible in several directions.

First, some settings were not previously envisaged, such as a model with very long lags of a variable when only a few lags matter. When one, or a few, important effects are hidden in a morass of irrelevance,

the pre-search block tests need not be appropriate, so we consider the outcome of a maximum t-test as a check (sub-section 2.1). We also use much looser significance levels for the block tests than in Hendry and Krolzig (1999), where the overall procedure was under-sized under the null.

Secondly, the calibration of the mis-specification heteroscedasticity tests was poor in Hendry and Krolzig (2003b), but this transpires to be a problem with the degrees of freedom assumed for the reference distribution (sub-section 2.2). The changed degrees of freedom lead to a substantial improvement in behaviour under the null.

Fnally, lag-order determination uses a combined top-down/bottom-up approach, complemented by an automatic Lagrange-multiplier test for omitted regressors. We also investigated exploiting the information in the ordered t-statistics in the GUM to locate a cut-off between included and excluded variables, but while suitable for orthogonal problems, multi-path search remains necessary in general: section 6 briefly addresses the collinearity issue. The post-selection procedure (14) is discussed in Krolzig and Hendry (2003).

2.1 Max t-tests

When only one of a large set n of candidate variables matters, then on average, a block test F^n_{T-n} will have low power to detect its relevance compared to a focused t-test. A crude approximation relating these two statistics, valid for orthogonal variables, is:

$$\mathsf{F}^n_{T-n} \simeq \frac{1}{n}\sum_{i=1}^{n} \mathsf{t}^2_{(i)}.$$

The expected value of $\mathsf{t}^2_{(i)}$ under the null is unity, so if $n-1$ variables are irrelevant, then on average, ignoring sampling variation:

$$\mathsf{F}^n_{T-n} \simeq \frac{1}{n}\sum_{i=1}^{n-1} 1 + \frac{1}{n}\mathsf{t}^2_{(n)} = 1 + \frac{1}{n}\left(\mathsf{t}^2_{(n)} - 1\right), \tag{1}$$

since $\mathsf{E}[\mathsf{t}^2_{(i)}|\mathsf{H}_0] = 1$, where $\mathsf{t}^2_{(n)}$ denotes the largest statistic. Let the block test be conducted at size α, then a $\max\{|\mathsf{t}|\}$ criterion with the correct size would use the approximate nominal significance level (see e.g., Savin, 1984):

$$\delta^{\alpha}_n = 1 - (1-\alpha)^{1/n}. \tag{2}$$

For example, for $n = 10$ when $\alpha = 0.05$ so $\mathsf{P}\big(\mathsf{F}^{10}_{90} > 1.935|\mathsf{H}_0\big) = 0.05$, then from (1), a significant outcome due to only $\mathsf{t}^2_{(10)}$ requires its value to be about 10.3, whereas from (2):

$$\delta^{0.05}_{10} = 1 - (1-0.05)^{1/10} = 0.0051,$$

which entails $\mathsf{t}^2_{(10)} > 8.2$, and so is distinctly smaller.

To investigate the quality of the approximation in (2), we undertook a Monte Carlo experiment with n IID central $\mathsf{t}(\nu)$ random variates, where $\nu = 30$ is the degrees of freedom. In each of the $M = 100,000$ replications, we calculated the $\max\{|\mathsf{t}_1|, \ldots, |\mathsf{t}_n|\}$ of the n random variables, and compared the t-prob of its $1-\alpha$ quantiles to the prediction of the δ^{α}_n rule. Figure 1 plots δ^{α}_n for $\alpha = 0.01$ and 0.05 and compares these to the 0.95 and 0.99 quantiles of the associated t-probabilities. The results demonstrate the quality of the approximation.

Nevertheless, one relevant variable can easily hide in a set where the overall outcome is insignificant. Such situations create a potential for conflicting inference—*PcGets* would judge the variable

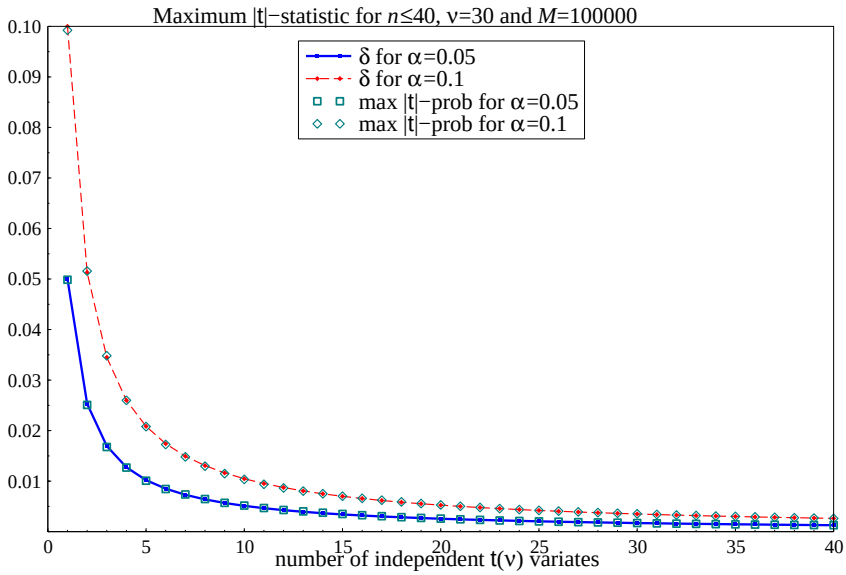

Figure 1 δ_n^α and $\max|\mathsf{t}|$ of n $\mathsf{IIDt}(\nu)$ random variates.

as irrelevant by the F test or a t-test based on δ_n^α, whereas a later investigator using a one-off t-test at significance level α would include it. Thus, we adopt a compromise between size and power which is more favourable to the latter when the initial specification is highly over-parameterized, but a few variables may matter, which is to consider the $\max\{|\mathsf{t}|\}$ statistic, but at a less stringent level than δ_n^α, namely twice the value from (2). In the above example, that would require $\mathsf{t}^2_{(10)} > 6.9$ (corresponding to the 1% level).

2.2 Recalibrating the heteroscedasticity tests

Krolzig and Hendry (2001) found that the QQ plots of the ARCH (see Engle, 1982) and unconditional heteroscedasticity tests (see White, 1980) were not straight lines, so the simulated outcomes did not match their anticipated distributions, and they therefore cautioned against their use. In reviewing *PcGets*, Dorian Owen (2003) suggested that the degrees of freedom were inappropriate by using a correction like that in Lagrange-multiplier autocorrelation tests (see e.g., Godfrey, 1978b, and Breusch and Pagan, 1980). Instead, as argued in (e.g.) Davidson and MacKinnon (1993, Ch. 11), since the covariance matrix is block diagonal between regression and scedastic function parameters, tests can take the former as given. Doing so, changes the statistics from being regarded as $\mathsf{F}_{\mathsf{arch}}(q, T-k-2q)$ and $\mathsf{F}_{\mathsf{het}}(q, T-k-q)$ to $\mathsf{F}_{\mathsf{arch}}(q, T-2q)$ and $\mathsf{F}_{\mathsf{het}}(q, T-q)$ respectively, and produces much closer matches with their anticipated distributions. Figure 2 shows the outcomes for all the mis-specification tests applied to the DGP, GUM and selected model in Krolzig and Hendry (2001).[2]

Overall, there is a marked improvement compared to the outcomes reported earlier.

3 Consistent selection

The performance of many selection algorithms as the sample size increases indefinitely is well known for an autoregressive process under stationarity and ergodicity: see Hannan and Quinn (1979) (whose criterion is denoted *HQ*), and Atkinson (1981), who proposes a general function from which various

[2]Chow 1 and 2 denote split-sample and forecast-period parameter constancy tests (see Chow, 1960); normality is the Doornik and Hansen (1994) test for normality; and AR denotes a 4th-order Lagrange multiplier test for residual autocorrelation: see Godfrey (1978a).

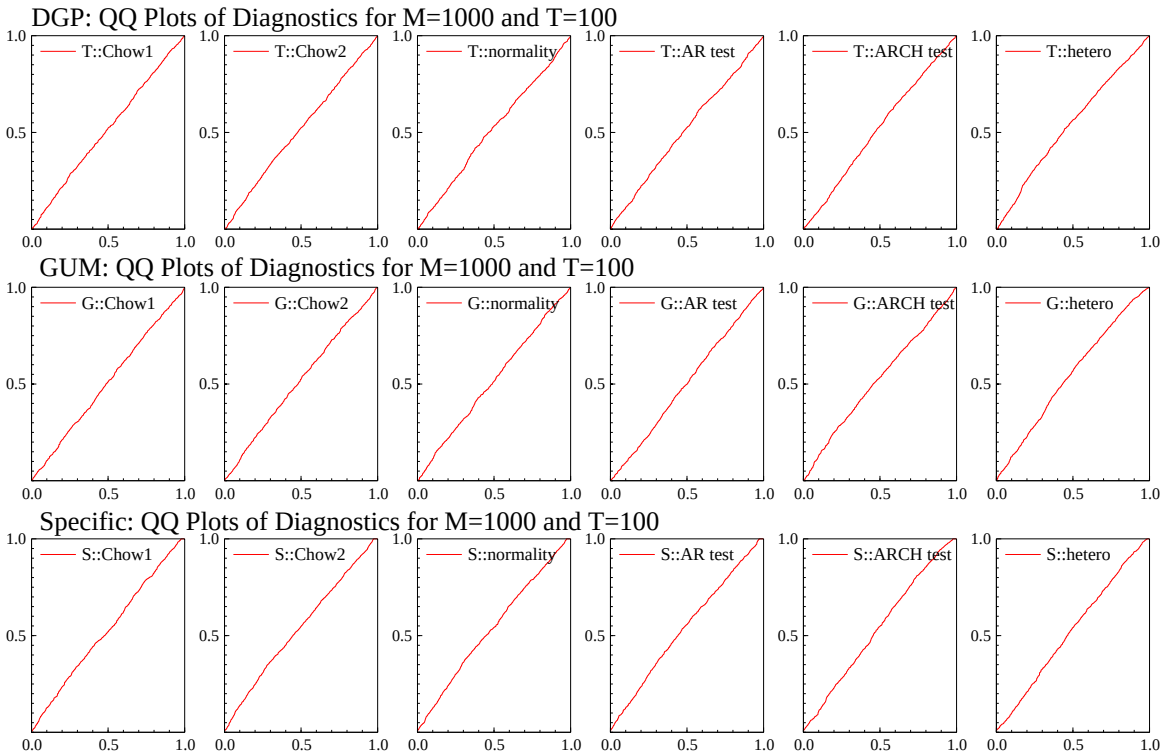

Figure 2 Mis-specification test operating characteristics.

criteria for model selection can be generated. The first criterion, proposed by Akaike (1969, 1973) (denoted *AIC* for Akaike information criterion) penalizes the log-likelihood by $2n/T$ for n parameters and a sample size of T, but does not guarantee a consistent selection as the sample size diverges. Both the Schwarz (1978) information criterion, also called the Bayesian information criterion, denoted *BIC*, and *HQ* are consistent, in that they ensure that a DGP nested within a model thereof will be selected with probability unity as T diverges relative to n. This requires that the number of observations per parameter diverges at an appropriate rate, so that non-centralities diverge (guaranteeing retention of relevant variables), and that the significance level decreases (so irrelevant variables are eventually almost surely not retained). In particular, *BIC* penalizes the log-likelihood by $n\log(T)/T$, whereas *HQ* uses $2n\log(\log(T))/T$, which Hannan and Quinn (1979) show is the minimum rate at which additional parameters must be penalized to ensure consistency. Then selection is strongly consistent when the assumed order of the model is no less than the true order and increases with the sample size. Based on a Monte Carlo, Hannan and Quinn (1979) suggest that *HQ* may perform better in large sample sizes.

PcGets implements similar requirements for consistent selection in both its Liberal and Conservative strategies. The general model must eventually be over-parameterized relative to the (local) DGP, and the nominal significance level decrease as the sample size increases. The Liberal strategy seeks to balance the chances of omitting variables that matter against retaining ones which are irrelevant in the DGP, so uses a relatively loose significance level (with HQ as its upper and BIC as its lower bound). The Conservative strategy uses a more stringent significance level in small samples, implicitly attributing a higher cost to retaining variables that do not matter in the DGP, but eventually converges on *BIC*. Figure 3 illustrates the *PcGets* rules for 10 variables relative to *AIC*, *BIC* and *HQ* for sample sizes up to 1000. As can be seen, the *PcGets* Conservative profile is much tighter than the three information criteria considered in small samples, whereas the Liberal strategy usually lies between *HQ* (as its upper bound)

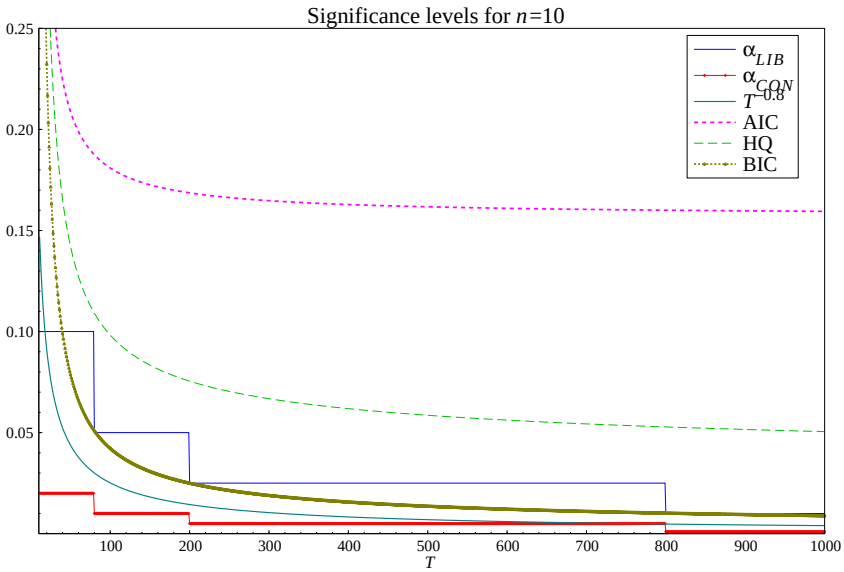

Figure 3 Significance level comparisons across selection rules.

and *BIC* (as its lower). The block jumps are those actually set for the two strategies over the range of sample sizes shown. A continuous profile could be implemented with ease, such as tracking *BIC* or one proportional to $T^{-0.8}$ (based on Hendry, 1995, Ch. 13: also shown), but as these strategies are designed for relatively non-expert users, it seems preferable to relate them to 'conventional' significance levels. The *AIC* is substantially less stringent, particularly at larger sample sizes, so would tend to over-select. However, the Conservative profile is noticeably tighter than *BIC* at small samples, so the next sub-section addresses its comparison with *BIC*. Importantly, while both *BIC* and *HQ* deliver consistent selections, they could differ substantively in small samples, which is precisely the intent of the two *PcGets* strategies. Thus, researchers must carefully evaluate the relative costs of over- versus under-selection for the problem at hand before deciding on the nominal significance level, or choice of strategy.

3.1 Comparisons with *BIC*

BIC selects from a set of n candidates the model with k regressors which minimizes:

$$BIC_k = \ln \widetilde{\sigma}_k^2 + c\frac{k \ln T}{T},$$

where $c \geq 1$ and:

$$\widetilde{\sigma}_k^2 = \frac{1}{T}\sum_{t=1}^{T}\left(y_t - \sum_{i=1}^{k}\widetilde{\beta}_i z_{i,t}\right)^2 = \frac{1}{T}\sum_{t=1}^{T}\widetilde{u}_t^2. \tag{3}$$

A full search for a fixed c and all $k \in [1, n]$ entails 2^n models to be compared, which for $n = 40$ exceeds 10^{12}. We focus on the implicit setting of significance levels involved in the choice of c (having shown in figure 3 the effect of altering the form of the penalty function), and the impact of pre-selection to reduce the value of n for a manageable number of models. First, we re-establish the formal link of *BIC* to significance levels.

Consider the impact of adding an extra orthogonalized regressor $z_{k+1,t}$, to a linear regression model with k such variables, so that:

$$\sum_{t=1}^{T} z_{k+1,t}\widetilde{u}_t = \sum_{t=1}^{T} z_{k+1,t}y_t - \sum_{t=1}^{T}\sum_{i=1}^{k}\widetilde{\beta}_i z_{i,t}z_{k+1,t} = \sum_{t=1}^{T} z_{k+1,t}y_t = \widehat{\beta}_{k+1}\sum_{t=1}^{T} z_{k+1,t}^2,$$

then, as is well known, from (3):

$$\begin{aligned}\widetilde{\sigma}^2_{k+1} &= \frac{1}{T}\sum_{t=1}^{T}\left(\widetilde{u}_t - \widehat{\beta}_{k+1}z_{k+1,t}\right)^2 = \widetilde{\sigma}^2_k\left(1 - \frac{\widehat{\beta}^2_{k+1}\sum_{t=1}^{T} z^2_{k+1,t}}{T\widetilde{\sigma}^2_k}\right)\\ &= \widetilde{\sigma}^2_k\left(1 - (T-k-1)^{-1}\widehat{\mathsf{t}}^2_{(k+1)}\frac{\widetilde{\sigma}^2_{k+1}}{\widetilde{\sigma}^2_k}\right),\end{aligned}$$

where:

$$\widehat{\mathsf{t}}^2_{(k+1)} = \frac{T\widehat{\beta}^2_{k+1}\sum_{t=1}^{T} z^2_{k+1,t}}{\widehat{\sigma}^2_{k+1}},$$

and:

$$\widehat{\sigma}^2_{k+1} = \frac{1}{T-k-1}\sum_{t=1}^{T}\widehat{u}^2_t \text{ for } \widehat{u}_t = \widetilde{u}_t - \widehat{\beta}_{k+1}z_{k+1,t}.$$

Consequently:

$$\widetilde{\sigma}^2_{k+1} = \widetilde{\sigma}^2_k\left(1 + (T-k-1)^{-1}\widehat{\mathsf{t}}^2_{(k+1)}\right)^{-1}, \tag{4}$$

so:

$$\begin{aligned}\mathit{BIC}_{k+1} &= \ln\widetilde{\sigma}^2_{k+1} + c\frac{(k+1)\ln T}{T}\\ &= \ln\widetilde{\sigma}^2_k + c\frac{k\ln T}{T} - \ln\left(1 + (T-k-1)^{-1}\widehat{\mathsf{t}}^2_{(k+1)}\right) + c\frac{\ln T}{T}\\ &= \mathit{BIC}_k + \frac{c}{T}\ln T - \ln\left(1 + (T-k-1)^{-1}\widehat{\mathsf{t}}^2_{(k+1)}\right).\end{aligned}$$

Hence, $\mathit{BIC}_{k+1} < \mathit{BIC}_k$ when:

$$\ln\left[T^{c/T}\left(1 + (T-k-1)^{-1}\widehat{\mathsf{t}}^2_{(k+1)}\right)^{-1}\right] < 0,$$

so the $(k+1)^{st}$ additional regressor will be retained by *BIC* when:

$$\widehat{\mathsf{t}}^2_{(k+1)} > (T-k-1)\left(T^{c/T} - 1\right).$$

Thus, choosing c is tantamount to choosing the p-value for the corresponding t-test. For example, when $T = 140$, with $c = 1$ (the usual choice), and $k = 40$, as in Hoover and Perez (1999), we have $\mathit{BIC}_{41} < \mathit{BIC}_{40}$ whenever $\widehat{\mathsf{t}}^2_{(41)} \geq 3.63$, or $|\mathsf{t}_{(41)}| \geq 1.9$.

To select no variables when the null model is true and $c = 1$ requires:

$$\widehat{\mathsf{t}}^2_{(k)} \leq (T-k)\left(T^{1/T} - 1\right) \ \forall k \leq n, \tag{5}$$

which is a sequence of $|\widehat{\mathsf{t}}_{(i)}|$ between 1.9 (at $k = 40$) and 2.24 (at $k = 1$) for $T = 140$. That outcome clearly entails at least every $|\widehat{\mathsf{t}}_{(i)}| < 1.9$ which has a probability, in an orthogonal setting, using even the best case 140 degrees of freedom as an approximation:

$$\mathsf{P}\left(|\mathsf{t}_{(i)}| < 1.9 \ \forall i = 1,\ldots,40\right) = (1 - 0.0595)^{40} \simeq 0.09. \tag{6}$$

Thus, 91% of the time, *BIC* should retain some variable(s). However, since there will be many 'highly insignificant' variables in a set of 40 irrelevant regressors, the bound of $|\widehat{\mathsf{t}}_{(i)}| < 2.2$ is probably the binding one, yielding (at the 'average' of 120 degrees of freedom), $\mathsf{P}\left(|\mathsf{t}_{(i)}| < 2.2 \ \forall i\right) \simeq 0.3$. Reducing

both T and k can worsen the chances of correct selection: for example, $T = 80$, $c = 1$ and $k = 30$ leads to a range between $\mathsf{P}\left(|\mathsf{t}_{(i)}| < 1.68\ \forall i = 1, \ldots, 30\right) \simeq 0.04$ and $\mathsf{P}\left(|\mathsf{t}_{(i)}| < 2.11\ \forall i\right) \simeq 0.31$. Such probabilities of correctly selecting a null model at relevant sample sizes are too low to provide a useful practical basis. Consequently, two amendments have been proposed.

First, lowering the maximum size of model to be considered using 'pre-selection' as in (say) Hansen (1999). He enforces a maximum of 10 in the *BIC* formula when $T = 140$ despite $n = 40$ by sequentially eliminating variables with the smallest t-values until 30 are removed. However, such a procedure entails that *BIC* actually confronts a different problem. If pre-selection did not matter, then under the null, we would have:

$$\mathsf{P}\left(|\mathsf{t}_{(i)}| < 2.16\ \forall i = 1, \ldots, 10\right) = (1 - 0.0325)^{10} = 0.72. \tag{7}$$

But the un-eliminated variables are those selected to have the largest t-values, so (7) overstates the performance of his approach. Conversely, (6) will understate what happens after pre-selection, because the very act of altering n changes the *parameters* of *BIC*, and is not just a 'numerical implementation'. Hansen in fact reports 0.45 for his Monte Carlo applied to the Hoover–Perez experiments. Interestingly, using the 'baseline' t-value of 1.9 in (6) yields:

$$\mathsf{P}\left(|\mathsf{t}_{(i)}| < 1.9\ \forall i = 1, \ldots, 10\right) = 0.54, \tag{8}$$

so even allowing for the initial existence of 40 variables matters considerably. A formal analysis requires calculating the conditional probability of the 10 largest t-values being insignificant at the critical value entailed by (5) for $n = 10$, given that the smallest 30 t-values have been excluded irrespective of their significance. Campos (2003) reports calculations for the setting where the smallest 30 are in fact insignificant, and finds 0.29 in place of the approximation of 0.54 in (8) or the unconditional 0.09 in (7).

Conversely, to have a higher chance of selecting the null model, one could increase c. For example, $c = 2$ raises the required $|\widehat{\mathsf{t}}_{(i)}|$ to 2.7 and:

$$\mathsf{P}\left(|\mathsf{t}_{(i)}| < 2.7\ \forall i = 1, \ldots, 40\right) = (1 - 0.0078)^{40} = 0.73, \tag{9}$$

which is a dramatic improvement over (6). Hansen's setting of $c = 2$ when $n = 10$ raises the required $|\widehat{\mathsf{t}}_{(i)}| < 3.08$, and again ignoring pre-selection, delivers a 97.5% chance of correctly finding a null model (he reports 95% in his Monte Carlo, whereas $(1 - 0.0078)^{10} = 0.92$, and Campos finds 0.85 for the conditional probability).

Nevertheless, when the null is false, both steps (i.e., raising c and/or arbitrarily simplifying till 10 variables remain) could greatly reduce the probability of retaining relevant regressors with t-values less than 2.5 in small samples. This effect does not show up in the Hoover–Perez experiments because the 'population' t-values are either very large or very small. Moreover, there are very few relevant variables whereas more than 10 would ensure an inconsistent selection.

Three conclusions emerge from this analysis. First, pre-selection can help locate the DGP by altering the 'parameters' entered into the *BIC* calculations, specifically the apparent degrees of freedom and the implicitly required t-value. *PcGets* employs a similar 'pre-selection' first stage, but based on block sequential tests with very loose significance levels so relevant variables are unlikely to be eliminated. Secondly, the trade-off between retaining irrelevant and losing relevant variables remains, and is determined by the choice of c implicitly altering the significance level. In problems with many t-values around 2 or 3, high values of c will be very detrimental. Thirdly, the asymptotic comfort of consistent selection when the model nests the DGP does not greatly restrict the choice of strategy in small samples. We also note that *BIC* does not address the difficulty that the initial model specification may not be adequate to characterize the data, but will still select a 'best' representation without evidence on how poor it may

be. In contrast, *PcGets* commences by testing for congruency: perversely, in Monte Carlo experiments conducted to date, where the DGP is a special case of the general model, such testing will lower the relative success rate of *PcGets*. Finally, the arbitrary specification of an upper bound on n is both counter to the spirit of *BIC*, and would deliver adverse findings in any setting where n was set lower than the number of relevant DGP variables.

4 Small-sample behaviour of *PcGets*

Table 1 summarizes the main features of the various Monte Carlo experiments conducted to date, and referred to below (HP, JEDC, S0–S4 and S0*–S4* respectively denote Hoover and Perez, 1999, Krolzig and Hendry, 2001, and two variants of the *PcGets* calibration experiments in Hendry and Krolzig, 2003b). We now summarize the operating characteristics of *PcGets* across the experiments in Table 1.

Table 1 Monte Carlo designs.

Design	regressors	causal	nuisance	$\|t\|$-values	avg. $\|t\|$-value
HP0	41	0	41		0
HP2*	41	1	40	5.77	5.77
HP2	41	1	40	11.34	11.34
HP7	41	3	38	(10.9, 16.7, 8.2)	11.93
JEDC	22	5	17	(2,3,4,6,8)	4.6
S0	34	0	34		0
S2	34	8	26	(2,2,2,2,2,2,2,2)	2
S3	34	8	26	(3,3,3,3,3,3,3,3)	3
S4	34	8	26	(4,4,4,4,4,4,4,4)	4
S0*	42	0	42		0
S2*	42	8	34	(2,2,2,2,2,2,2,2)	2
S3*	42	8	34	(3,3,3,3,3,3,3,3)	3
S4*	42	8	34	(4,4,4,4,4,4,4,4)	4

First, we consider control over 'size', such that the actual null rejection frequencies are close to the nominal levels set by the user, 'independently' of the problem investigated. Figure 4 plots the ratio of actual to nominal size across the various studies re-analyzing the Lovell (1983) (aka Hoover–Perez) experiments at 5% and 1% nominal levels. The outcomes confirm that stabilization has occurred as we have learned more about how such algorithms function, and so improve their search procedures. The most recent size estimates incorporate the sub-sample reliability weightings, and are slightly below nominal—despite there being between 35 and 40 irrelevant regressors. Consequently, 'overfitting' in the sense of finding too many significant variables does not occur, especially as such large numbers of irrelevant variables are not representative of empirical problems.

Secondly, we consider the calibration accuracy of the two basic strategies, Conservative and Liberal. Figure 5 graphically illustrates four main aspects of the outcomes across all the Monte Carlo experiments to date for both strategies. Panel a concerns a different sense of 'overfitting', namely potential downward biased estimates of the equation standard error, $\widehat{\sigma}$, for the true value σ. Again this does not occur: the final average $\widehat{\sigma}$ is close to σ. The Liberal strategy has a slight downward bias (less than 5% of σ), whereas the Conservative is upward biased by a similar amount. Such behaviour is easily explained: the latter eliminates variables which matter so fits worse than the GUM, which unbiasedly estimates σ, and the former retains some variables which only matter by chance, but thereby slightly over fits. It must be stressed that *PcGets* model selection is not based on fit as a criterion, but a minimal congruent

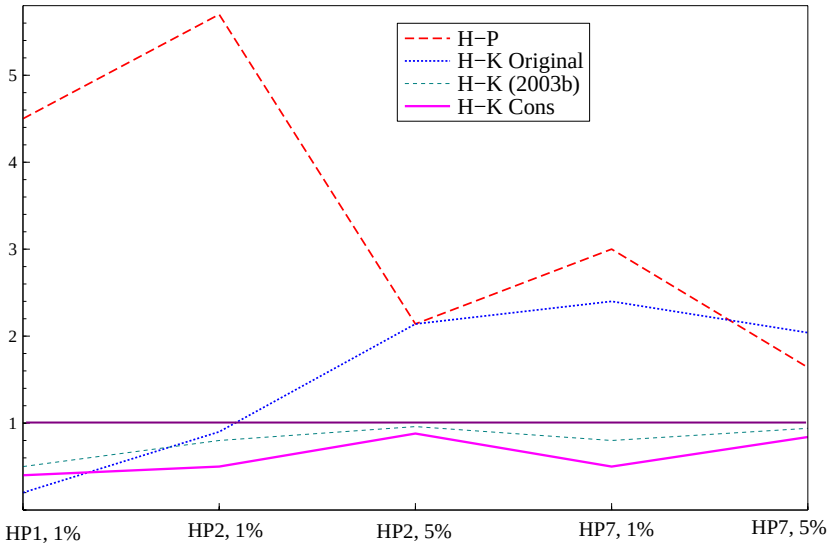

Figure 4 Ratios of actual to nominal sizes in the data-mining experiments.

encompassing model will necessarily have the best fit at the chosen significance level. Equation (4) recorded the fit relationship between models of size k and $k+1$, which can be re-expressed for unbiased estimators of σ (i.e., corrected for degrees of freedom) as:

$$\frac{\widehat{\sigma}_k^2}{\widehat{\sigma}_{k+1}^2} = 1 + \frac{\widehat{\mathsf{t}}_{(k+1)}^2 - 1}{T-k}. \tag{10}$$

The probability under the null that $|\mathsf{t}| > 2.5$ is 0.014 (when $T = 110$ and $k = 10$) so larger t-values will occur less than once in 70 draws under the null, yet the ratio in (10) would only be about 1.05.

Panel b shows sizes for the strategies across all experiments compared to their intended significance levels of 5% and 1%, both with and without sub-sample reliability weightings (denoted (rel) in the graphs): the latter are close to their targets, and in no case are deviations substantial for the former.

Panel c plots 'power', namely the average rejection frequency of the null for variables that do matter. The Conservative strategy naturally has no higher power than the Liberal, and reveals that the cost of avoiding spurious variables can be high in terms of missing variables that matter. The graphs also show the impact of the sub-sample reliability weightings on the resulting power, confirming that there is only a small effect, even at quite low powers where it should have most impact. Finally, comparisons between neighbouring successive Sj and Sj* experiments shows that the impact on power of 8 additional irrelevant variables is small, especially for the Liberal strategy.

Finally, figure 5d graphs the probabilities of locating the DGP, together with the corresponding outcomes when the search commences from the DGP, with tests conducted at 5% and 1%. The movements of the four lines are similar, and frequently the apparent problem for a search algorithm transpires to be a cost of inference since the DGP is sometimes never retained even when it is the initial specification. The out-performance of commencing from the DGP in the Hoover–Perez experiments is owing to the high degree of over-parameterization and very large t-values on relevant variables, but even so, the Conservative strategy does a respectable job. When population t-values are 2 or 3, the Liberal strategy does best, and sometimes outperforms commencing from the DGP with a 1% significance level (S3 and S4). Notice also that the two strategies cannot be ranked on this criterion: their relative performance depends on the unknown state of nature. Nevertheless, as Hendry and Krolzig (2001, Ch. 5) discuss, a user may be aware of the 'type' of problem being confronted, in which case, figure 5d shows the potential advantages of an appropriate choice of strategy combined with a good initial specification.

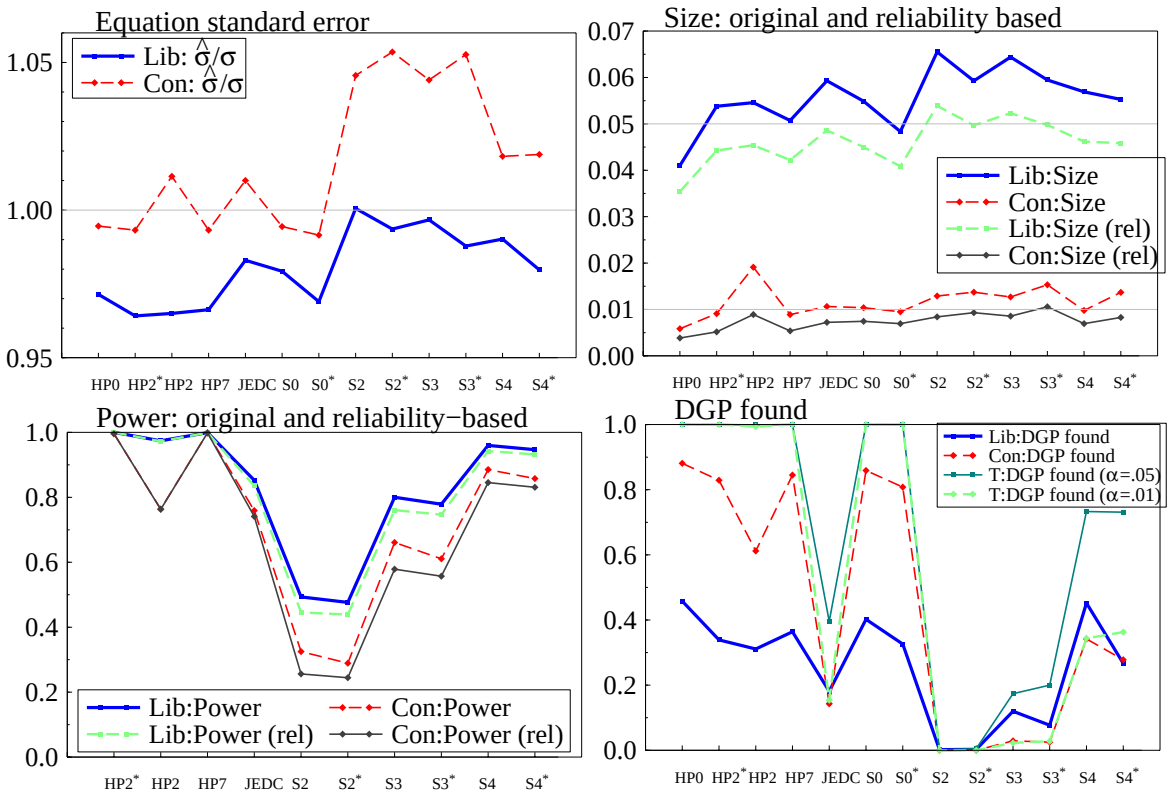

Figure 5 Overview of accuracy, size, power and success.

These findings also confirm the closeness in practice of the strategies to their desired operating characteristics.

5 'Pre-test' and 'selection' effects in small samples

Statistical tests have non-degenerate null distributions, and hence have non-zero size, and (generally) non-unit power. Consequently, even if the local DGP were correctly specified *a priori* from economic theory, when an investigator did not know that the resulting model was 'true' – so sought to test hypotheses about its coefficients – then inferential mistakes could occur, the seriousness of which depend on the characteristics of the local DGP and the sample drawn. Should the selected model thereby differ from the DGP, it will deliver biased coefficient estimates: this is called the 'pre-test' problem, since unbiased estimates could have been obtained from the unrestricted model by conducting no selection tests (see e.g., Judge and Bock, 1978). The arguments against using alternatives such as Stein-James 'shrinkage' are presented in Hendry and Krolzig (2003b). Assuming that one knows the truth, and knows that one does, so no testing is needed, is not a relevant benchmark in economics. In the following simulations, the role of selections commencing from the DGP is merely to measure the additional costs of selection compared to commencing from the GUM.

5.1 Selection effects on coefficient estimates

To investigate the impact of selection, we re-ran the Krolzig and Hendry (2001) experiments. As shown in table 2, unconditionally coefficient estimates are downward biased (being a mix of 0 and $\widehat{\beta}_i$ when z_i is retained). However, the Liberal strategy biases are under 10% for $|\mathtt{t}| > 3$.

Table 2 Unconditional coefficient estimates, SEs and SDs (including zeros).

	DGP	Reduction of DGP		GUM	Reduction of GUM		true value
variable		LIB	CON		LIB	CON	
mean							
Z_a	0.204	0.142	0.092	0.204	0.140	0.093	0.200
Z_b	0.301	0.270	0.230	0.300	0.271	0.224	0.300
Z_c	0.399	0.396	0.378	0.399	0.393	0.373	0.400
Z_d	0.604	0.601	0.601	0.604	0.604	0.604	0.600
Z_e	0.803	0.796	0.796	0.801	0.803	0.803	0.800
SE							
Z_a	0.103	0.051	0.029	0.113	0.049	0.029	0.100
Z_b	0.102	0.083	0.066	0.112	0.081	0.063	0.100
Z_c	0.103	0.100	0.093	0.113	0.097	0.091	0.100
Z_d	0.102	0.103	0.104	0.113	0.101	0.103	0.100
Z_e	0.103	0.103	0.103	0.113	0.101	0.103	0.100
SD							
Z_a	0.103	0.150	0.150	0.115	0.151	0.150	
Z_b	0.102	0.149	0.182	0.113	0.150	0.184	
Z_c	0.103	0.113	0.151	0.115	0.125	0.158	
Z_d	0.103	0.104	0.107	0.116	0.108	0.108	
Z_e	0.106	0.100	0.102	0.119	0.111	0.110	
residuals							
σ	0.998	1.007	1.017	0.998	0.981	1.008	1.000
% bias	-0.2%	0.7%	1.7%	-0.2%	-1.9%	0.8%	

Table 3 Conditional coefficient estimates, SEs and SDs (excluding zeros).

	DGP	Reduction of DGP		GUM	Reduction of GUM		true value
variable		LIB	CON		LIB	CON	
mean							
Z_a	0.204	0.286	0.324	0.204	0.285	0.322	0.200
Z_b	0.301	0.332	0.358	0.300	0.333	0.360	0.300
Z_c	0.399	0.407	0.420	0.399	0.410	0.422	0.400
Z_d	0.604	0.602	0.602	0.604	0.604	0.605	0.600
Z_e	0.803	0.796	0.796	0.801	0.803	0.803	0.800
SE							
Z_a	0.103	0.102	0.101	0.113	0.099	0.101	0.100
Z_b	0.102	0.102	0.102	0.112	0.100	0.100	0.100
Z_c	0.103	0.103	0.103	0.113	0.101	0.102	0.100
Z_d	0.102	0.103	0.104	0.113	0.101	0.103	0.100
Z_e	0.103	0.103	0.103	0.113	0.101	0.103	0.100
SD							
Z_a	0.103	0.066	0.061	0.115	0.070	0.062	
Z_b	0.102	0.082	0.075	0.113	0.084	0.075	
Z_c	0.103	0.095	0.089	0.115	0.098	0.090	
Z_d	0.103	0.102	0.104	0.116	0.108	0.106	
Z_e	0.106	0.100	0.102	0.119	0.111	0.110	
residuals							
σ	0.998	1.007	1.017	0.998	0.981	1.008	1.000
% bias	-0.2%	0.7%	1.7%	-0.2%	-1.9%	0.8%	

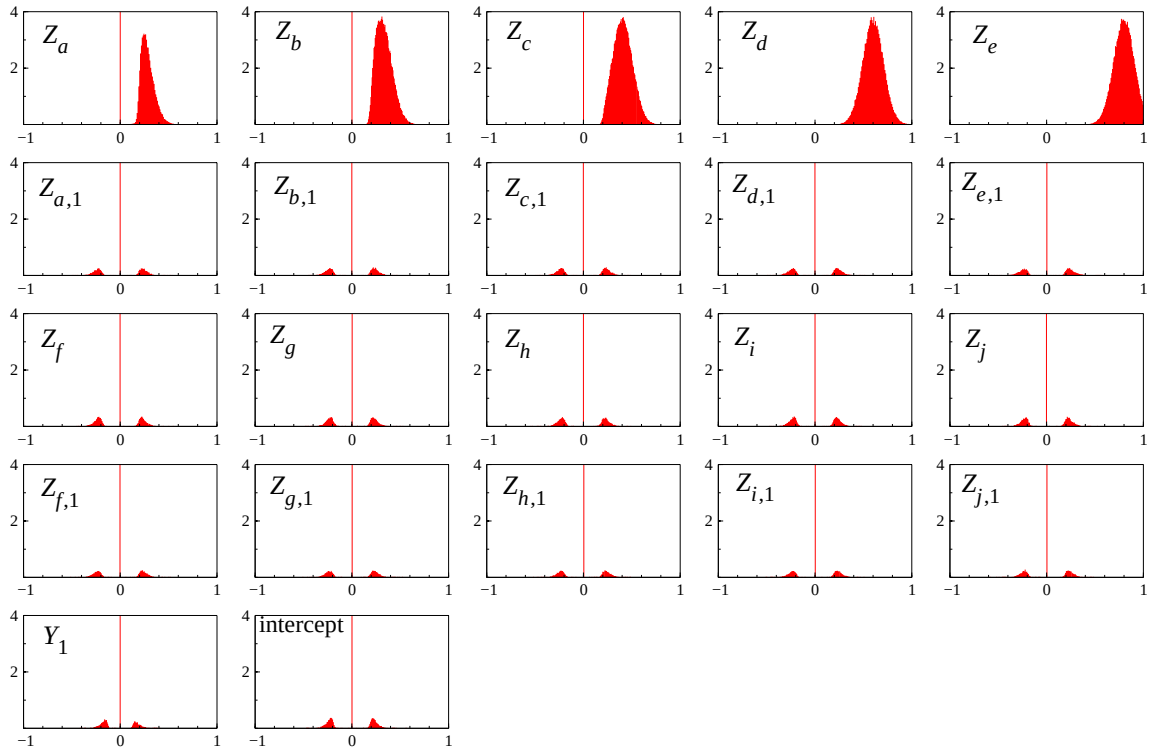

Figure 6 Unconditional distributions from the Liberal strategy.

Figure 6 shows the unconditional distributions of the five relevant and 17 irrelevant regressors for the Liberal strategy.[3] These unconditional distributions illustrate the quality of the classification of variables into DGP variables (top row) and nuisance variables (all others). The non-zero-mass distribution of the DGP variables is truncated normal, but truncation does not affect variables with a population t-value greater than 4.

Conditional on being retained, the results are shown in table 3. As expected, the coefficient estimates are now upward biased for smaller t-values ($|t| \leq 3$), more so for the Conservative strategy, but are close to the population values for larger t-values. The Liberal strategy biases are under 10% for $|t| > 3$.

Figure 7 records the corresponding conditional distributions. Those for the non-DGP variables are bimodal and symmetric, except for the lagged endogenous variable, where the impact of the famous Hurwicz (1950) bias is clear.

The final important result is that these 'pre-test' effects are not, in any essential respects, changed by search *per se*. The coefficient biases are closely similar when commencing from the DGP and the GUM for each strategy, both conditionally and unconditionally as tables 2 and 3 show.

5.2 Selection effects on estimated standard errors and standard deviations

Crucially, the conditional estimated standard errors (SEs) are not biased on either strategy, so the reported SEs for a selected equation's coefficients are close to providing unbiased estimates of the standard deviations (SDs) for the estimated DGP. At first sight, that might seem an astonishing result, since despite selection, the estimated uncertainty when a DGP variable is selected is a correct reflection of the

[3]The results for the Conservative strategy are similar, but distributions of irrelevant variables are almost invisible, and so are not shown.

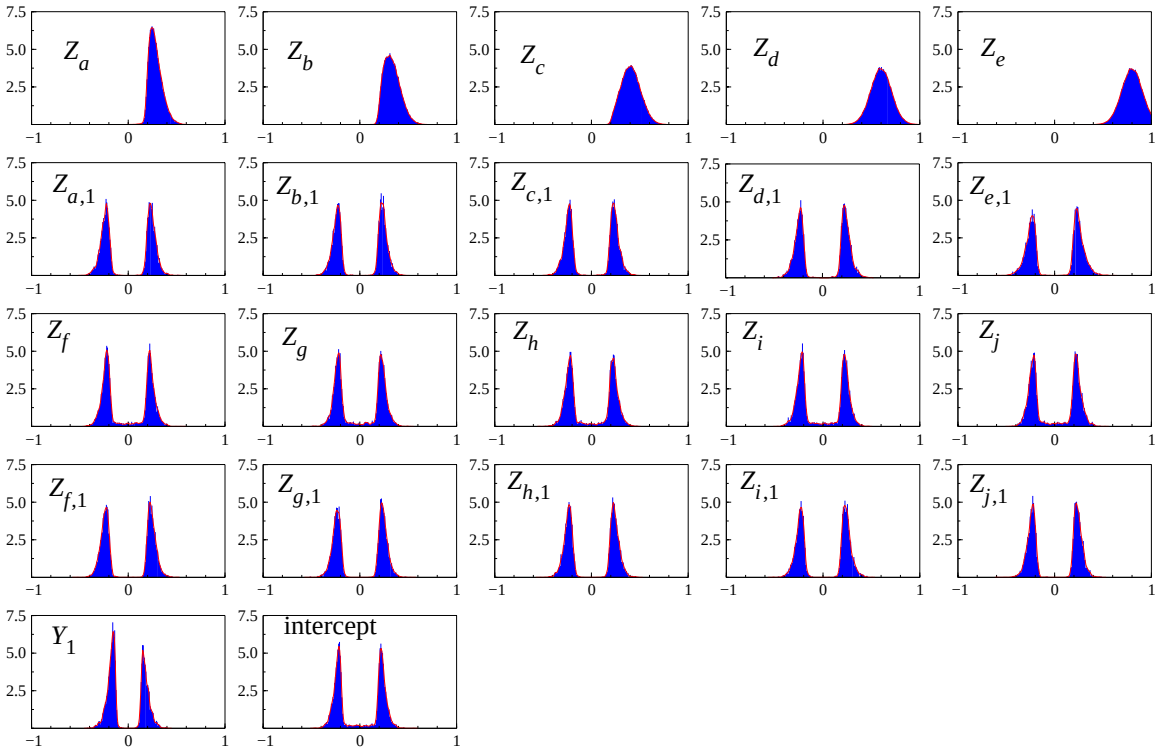

Figure 7 Conditional distributions from the Liberal strategy.

uncertainty operating in the DGP without selection. However, the intuition is simple: the SDs in the estimated DGP model are correctly estimated by the reported SEs (column 2); the latter are based on the estimated equation standard error ($\widehat{\sigma}$, which is close to σ on average as shown on the bottom row) times the associated square-root element from $(\mathbf{X}'\mathbf{X})^{-1}$; and that in turn is approximately the same in the selected model when the relevant variable is retained.

Naturally, unconditional SEs are downwards biased (as parameter estimates restricted to zero have zero standard errors), and the SDs are upward biased (again as a mix of 0 and $\widehat{\beta}_i$). The probability p of retaining a variable with a population t^2-value of 4 is approximately 0.5, so the effects are largest at small population t-values. Indeed, the mean unconditional estimates and their SEs are approximately p times the corresponding conditional.

However, the relevance of such unconditional 'sampling properties' is unclear in the context of model selection when the DGP is unknown. The elimination of insignificant variables is the objective of simplification in small samples, and the underlying state of nature (whether variables are relevant or irrelevant) is unknown, so the cost of the bimodality of the unconditional selection distribution for relevant variables is a larger SD.

As noted earlier, in almost all cases, the estimated equation standard errors are close to σ, so that *PcGets* does not 'overfit'. Rather, the Conservative strategy underfits by eliminating too many of the relevant regressors in its attempt to avoid adventitious significance, whereas the Liberal strategy performance depends on the number of irrelevant variables in the GUM, and can be either under or over σ. Indeed, so can the SEs and SDs, both conditional on retaining a variable, and unconditionally.

Overall, these results confirm using the Liberal strategy as the default option.

5.3 Selection effects on the two heteroscedasticity tests

Another feature of interest is the impact of model selection on the outcomes of test statistics. This is shown in figure 8 for the two heteroscedasticty tests recalibrated in section 2.2. The graphs compare the ratios of actual sizes to nominal in the DGP, GUM and the selected model.[4]

The operational rules adopted were as follows. Specific models with diagnostic tests indicating an invalid reduction at 1% or less were rejected if the GUM showed no mis-specifications at 5%. If a mis-specification test was significant at 1%, the test was dropped from the test battery. If the p-value of the mis-specification test was between 1% and 5%, the significance level was reduced from 1% to 0.5%.

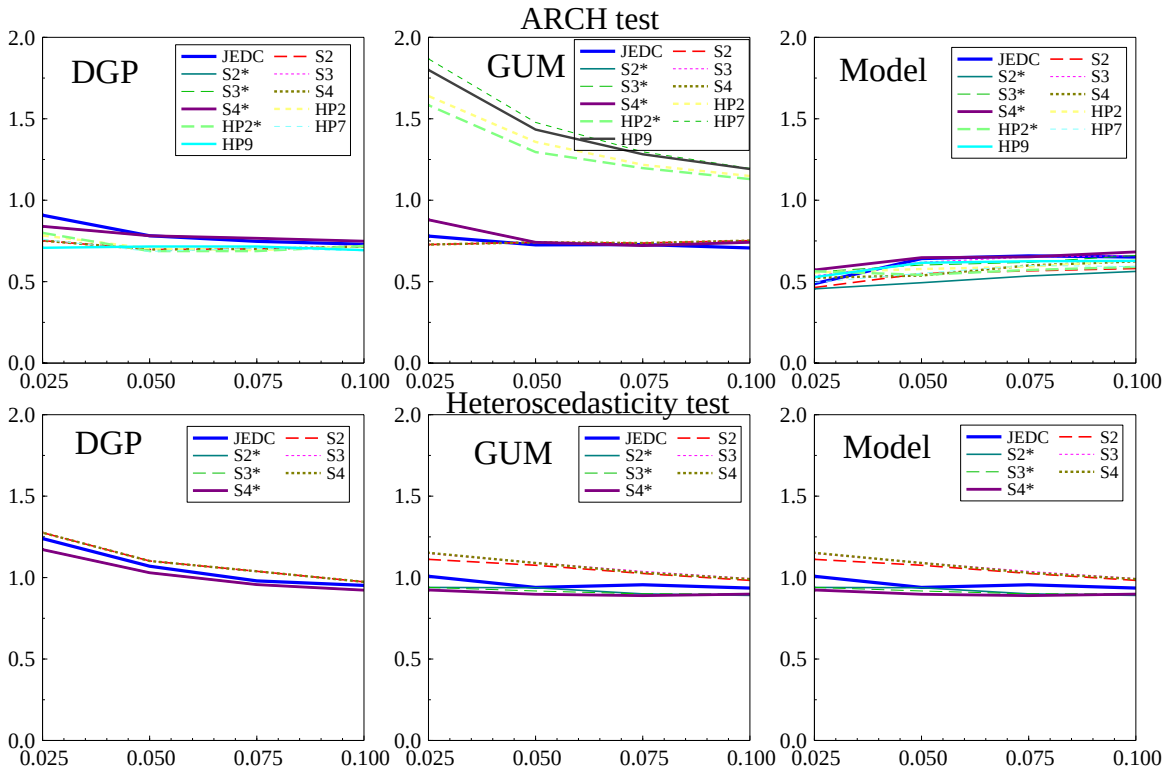

Figure 8 Ratios of actual sizes to nominal in the DGP, GUM and selected model.

There is almost no change in the rejection frequencies for quantiles above the nominal significance level, but an increasing impact as the quantile decreases. The latter effect is essentially bound to occur, since models with significant heteroscedasticity are selected against by construction. Nevertheless, the outcomes in these graphs do not represent a 'distortion' of the sampling properties: the key decision is taken at the level of the general model, and conditional on not rejecting there, no change should occur in that decision. At most nominal significance levels in the GUM, the tests have their anticipated operating characteristics, with the ARCH test oversized at smaller significance levels in the HP experiments due to the heteroscedastic nuisance regressors affecting the residuals.

[4]The 1% level showed larger departures, but was imprecisely estimated given the rarity with which it occurred, and has been omitted from the graphs.

6 Collinearity

Perfect collinearity denotes an exact linear dependence between variables; perfect orthogonality denotes no linear dependencies; but any intermediate state depends on which 'version' of a model is inspected, as collinearity is not invariant under linear transforms. *PcGets* provides a 'collinearity analysis', reporting the correlation matrix and its eigenvalues, but because of a lack of invariance, suitable statistics are unclear. First, eigenvalues are only invariant under orthogonal, and not under linear, transforms, so depend on the transformations of the variables (rather than the 'information content'). Secondly, observed correlations are not reliable indicators of potential problems in determining if either or both of two variables should enter a model – the source of their correlation matters. For example, inter-variable correlations above 0.999 can easily arise in systems with unit roots and drift, but there is little difficulty determining the relevance of variables. For example, when the DGP is:

$$\begin{aligned} y_t &= \alpha_0 + \alpha_1 y_{t-1} + \alpha_2 z_t + \epsilon_t \text{ with } \epsilon_t \sim \mathsf{IN}\left[0, \sigma_\epsilon^2\right] \\ z_t &= \gamma + z_{t-1} + v_t \text{ with } v_t \sim \mathsf{IN}\left[0, \sigma_v^2\right], \end{aligned} \tag{11}$$

where $\mathsf{E}\left[\epsilon_t v_s\right] = 0 \; \forall t, s$ and the fitted model is (say):

$$y_t = \beta_0 + \beta_1 y_{t-1} + \beta_2 z_t + \beta_3 z_{t-1} + \cdots + u_t,$$

even if all correlations exceed 0.999, neither Liberal nor Conservative strategy have great difficulty retaining z_t. In essence, the model is isomorphic to:

$$\Delta y_t = \alpha_0 + \alpha_2 \Delta z_t + (\alpha_1 - 1)\left(y_{t-1} - \kappa z_{t-1}\right) + \epsilon_t$$

where Δz_t and $(y_{t-1} - \kappa z_{t-1})$ are little correlated.

Conversely, in a bivariate normal:

$$\begin{pmatrix} x_t \\ z_t \end{pmatrix} \sim \mathsf{IN}_2\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right], \tag{12}$$

with a conditional model as the DGP:

$$y_t = \beta x_t + \gamma z_t + \epsilon_t \tag{13}$$

when $\rho = 0.99$ there is almost no hope of determining which variables matter in (13).

Transforming variables to a 'near orthogonal' representation before modelling can help resolve this problem, but otherwise, eliminating one of the two variables seems inevitable. Which is dropped depends on the vagaries of sampling, inducing considerable 'model uncertainty', as the selected model oscillates between retaining x_t or z_t (or both): either variable is an excellent proxy for the dependence of y_t on $\beta x_t + \gamma z_t$. That remains true even when one of the variables is irrelevant, although then the multiple-path search is likely to select the correct equation. When both are relevant, a Monte Carlo model-selection study of (13) given (12) when $\rho = 0.99$ would almost certainly suggest that the algorithm had a low probability of selecting the DGP. In empirical applications, however, for users willing to carefully peruse the detailed output, the impact of collinearity will be manifest in the number of different terminal models entered in encompassing comparisons. Such information could guide selection when subject-matter knowledge was available.

A serious indirect cost imposed by collinearity is that the t-values in the GUM are poor indicators of the importance of variables. Thus, tests which use the initial ordered $\mathsf{t}^2_{(i)}$ as a guide to the selection

of candidate variables for elimination cannot perform adequately, which includes the initial cumulative F-test and block tests (e.g., on groups of lagged variables). Thus, a simple separation into 'included' and 'excluded' variables in a one-off test is infeasible under non-orthogonality, and multi-path searches are essential. Transforming the variables to a 'near orthogonal' representation before modelling probably requires analyzing the properties of the regressors, and takes us in the direction of a system variant of *Gets*: for applications of such ideas in the context of a vector autoregression, see Krolzig (2000).

The effects of collinearity on the selection properties of *PcGets* are illustrated by a variation of the Monte Carlo experiments in Krolzig and Hendry (2001), The DGP is a Gaussian regression model, where the strongly-exogenous variables are independent Gaussian AR(1) processes:

$$\begin{array}{rclrcl} y_t & = & \sum_{k=1}^{5} \beta_{k,0} z_{k,t} + u_t, & u_t & \sim & \mathsf{IN}\left[0, \sigma_u\right], \\ z_t & = & (\alpha \mathbf{I}_{10}) z_{t-1} + v_t, & v_t & \sim & \mathsf{IN}_{10}\left[\mathbf{0}, \left(1-\alpha^2\right)\sigma_v^2 \mathbf{I}_{10}\right] \text{ for } t = 1, \ldots, T. \end{array} \tag{14}$$

The parameterization of the DGP is $\beta_{1,0} = 0.2$, $\beta_{2,0} = 0.3$, $\beta_{3,0} = 0.4$, $\beta_{4,0} = 0.6$, $\beta_{5,0} = 0.8$, and $\sigma_u^2 = \sigma_v^2 = 1$. The population t-value associated with regressor k is given by:

$$\mathsf{t}_k = \beta_k \sqrt{T} \frac{\sigma_z}{\sigma_u} = \beta_k \sqrt{T} \frac{\left(1-\alpha^2\right)\sigma_v}{(1-\alpha^2)\sigma_u} = \beta_k \sqrt{T}$$

The DGP is designed to ensure invariant population t-values with increasing α. For $T = 100$, the non-zero population t-values are therefore $2, 3, 4, 6, 8$.

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 $z_{6,t}, \ldots, z_{10,t}$ and the first lags of all regressors:

$$y_t = \pi_{0,0} + \pi_{0,1} y_{t-1} + \sum_{k=1}^{10} \sum_{i=0}^{1} \pi_{k,i} z_{k,t-i} + w_t, \;\; w_t \sim \mathsf{IN}\left[0, \sigma_w^2\right]. \tag{15}$$

In an alternative experiment, we consider the orthogonal representation of (15):

$$y_t = \pi_{0,0} + \pi_{0,1} y_{t-1} + \sum_{k=1}^{10} \pi_k z_{k,t} + \sum_{k=1}^{10} \gamma_k \left(\alpha z_{k,t} - z_{k,t-1}\right) + w_t, \;\; w_t \sim \mathsf{IN}\left[0, \sigma_w^2\right]. \tag{16}$$

In (15) as in (16), 17 of 22 regressors are 'nuisance'. The sample size T is just 100, and the number of replications M is 1000. In a third experiment, using (16), the sample size is corrected for the time dependence of the regressors: $T(\alpha) = 100(1-\alpha^2)^{-1}$.

The Monte Carlo results are summarized in figure 9 which plots the size, power and the probability of finding the DGP with *PcGets* when commencing from (i) GUM (15) with $T = 100$, (ii) GUM (16) with $T = 100$, and (iii) GUM (16) with $T(\alpha)$. The first experiment illustrates the effects the collinearity: a significant loss in power and growing size. Starting from an orthogonalized GUM stabilizes size and power, which become α-invariant if the sample size is adjusted, although there is a slight fall in the probability of locating the DGP.

7 Conclusion

Model selection is an important part of a progressive research strategy, and itself is progressing rapidly. The automatic selection algorithm in *PcGets* provides a consistent selection like *BIC*, but in finite samples both ensures a congruent model and can out-perform in important special cases without *ad hoc* adjustments. Recent improvements have stabilized the size relative to the desired nominal significance

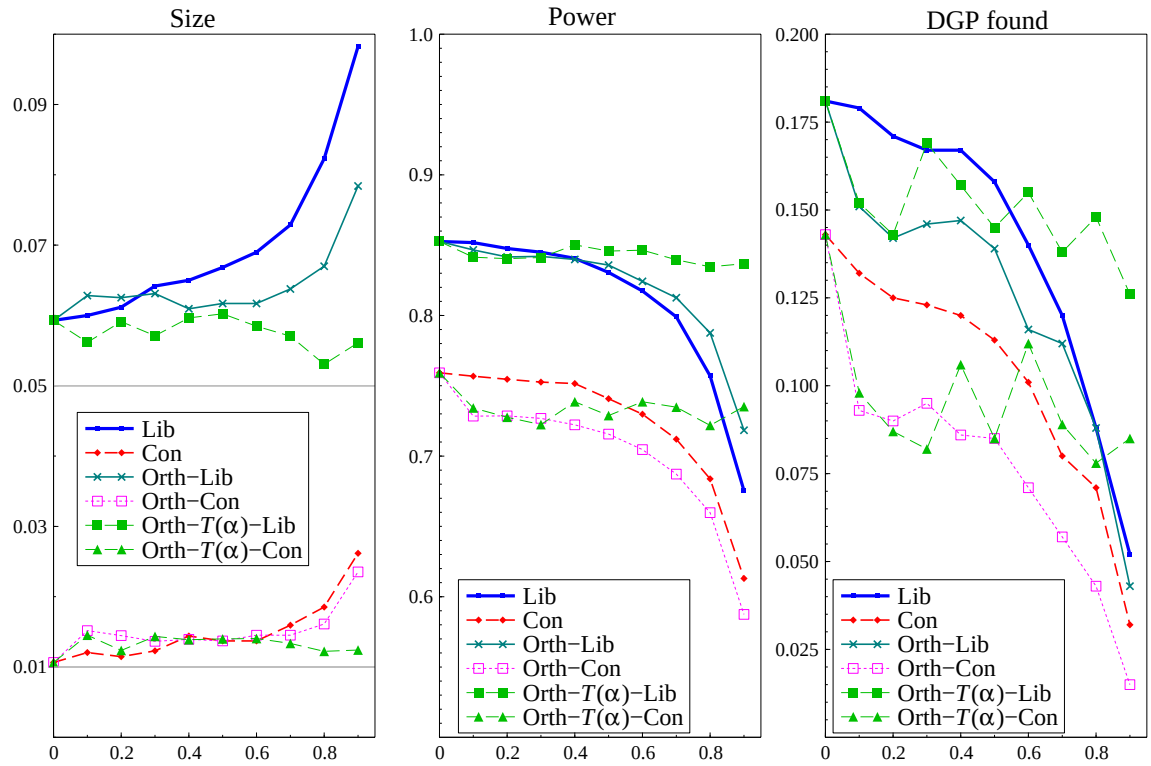

Figure 9 Selection properties of *PcGets* for varying α.

level, and the power relative to that feasible when the DGP is the initial specification. The power performance on recent Monte Carlo experiments is close to the upper bound of a scalar t-test at the given non-centrality from a known distribution, so the direction of improvement is to protect against specific formulations, such as needlessly long lags when a subset may matter.

Search *per se* does not seem to impose serious additional costs over those of inference (nor does the mis-specification testing, as that is needed even when commencing from the DGP specification). The results to date on 'pre-test' biases confirm that these arise from simplifying the DGP, not from searching for it in an over-parameterized representation. The equation standard error is found within $\pm 5\%$ of the population value, depending on the strategy adopted, so *PcGets* has no substantive tendency to 'overfit'. Depending on the state of nature, *PcGets* can even have a higher probability of finding the DGP starting from the GUM using the Liberal strategy, than a researcher commencing from the DGP but selecting by the Conservative strategy. Such a finding would have seemed astonishing in the aftermath of Lovell (1983), and both shows the progress achieved and serves to emphasize the importance of the choice of strategy for the underlying selection problem. That estimated standard errors in selected models are close to those that would be reported for sampling standard deviations in the estimated DGP might have surprised even more. The key to such performance seems to lie in using a search algorithm that commences from a congruent representation, explores feasible paths while retaining congruence, and terminates with a dominant encompassing selection.

Non-orthogonal designs remain problematic, and may be an area where expert knowledge will continue to prove very valuable. Nevertheless, we have added a 'quick modeller' option for non-expert users, which may be able to outperform all but expert econometricians in selecting from an initial dynamic GUM. The main difference from standard 'expert usage' is that the program chooses the maximum lag length in dynamic models, then checks the congruence of the resulting GUM before estimating

the levels representation unrestrictedly. The *PcGive* unit-root test (see e.g., Banerjee and Hendry, 1992, Ericsson and MacKinnon, 1999) is applied to check for possible cointegration, and if found, the program transforms the variables to differences and the cointegrated combination, then re-estimates that I(0) GUM, from which the usual procedures are applied to select a parsimonious undominated model.

What lies ahead? Certainly, the theoretical context assumed above of regression analysis 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 for endogenous regressors where sufficient valid instruments exist. Moreover, Omtzig (2002) has proposed an algorithm for automatic selection of cointegration vectors, and *Gets* is just as powerful a tool on cross-section problems, as demonstrated by Hoover and Perez (2000).

Selection with more candidate regressors than observations ($n > T$) is even feasible when the DGP is estimable (with $k < T/2$ regressors say), by successively following paths corresponding to all combinations of sub-blocks of initial variables and collecting all terminal models; then iterating from blocks of that set: see Hendry and Krolzig (2003a), who also apply that idea to selecting non-linear representations. Thus, we remain confident that further developments will continue to improve the performance of, and widen the scope of application for, automatic modelling procedures.

References

Akaike, H. (1969). Fitting autoregressive models for prediction. *Annals of the Institute of Statistical Mathematics*, **21**, 243–247.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In Petrov, B. N., and Csaki, F. (eds.), *Second International Symposion on Information Theory*. Budapest: Akademia.

Atkinson, A. C. (1981). Likelihood ratios, posterior odds and information criteria. *Journal of Econometrics*, *16*(1), 15–20.

Banerjee, A., and Hendry, D. F. (1992). Testing integration and cointegration: An overview. *Oxford Bulletin of Economics and Statistics*, **54**, 225–255.

Breusch, T. S., and Pagan, A. R. (1980). The Lagrange multiplier test and its applications to model specification in econometrics. *Review of Economic Studies*, **47**, 239–253.

Campos, J. (2003). Conditional probabilities in BIC after pre-selection. Mimeo, Economics Department, University of Salamanca.

Campos, J., Ericsson, N. R., and Hendry, D. F. (2003). Editors' introduction. In Campos, J., Ericsson, N. R., and Hendry, D. F. (eds.), *Readings on General-to-Specific Modeling*. Cheltenham: Edward Elgar. Forthcoming.

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

Davidson, R., and MacKinnon, J. G. (1993). *Estimation and Inference in Econometrics*. Oxford: Oxford University Press.

Doornik, J. A. (2001). *Object-Oriented Matrix Programming using Ox* 4th edn. London: Timberlake Consultants Press.

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 heteroscedasticity, with estimates of the variance of United Kingdom inflation. *Econometrica*, **50**, 987–1007.

Ericsson, N. R., and MacKinnon, J. G. (1999). Distributions of error correction tests for cointegration. International finance discussion paper no. 655, Federal Reserve Board of Governors, Washington, D.C. www.bog.frb.fed.us/pubs/ifdp/1999/655/default.htm.

Godfrey, L. G. (1978a). Testing against general autoregressive and moving average error models when the regressors include lagged dependent variables. *Econometrica*, **46**, 1293–1301.

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

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.

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

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., and Krolzig, H.-M. (2003a). Model selection with more variables than observations. Unpublished paper, Economics Department, Oxford University.

Hendry, D. F., and Krolzig, H.-M. (2003b). New developments in automatic general-to-specific modelling. In Stigum, B. P. (ed.), *Econometrics and the Philosophy of Economics*. Princeton: Princeton University Press. forthcoming.

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.

Hurwicz, L. (1950). Least squares bias in time series. In Koopmans, T. C. (ed.), *Statistical Inference in Dynamic Economic Models*, No. 10 in Cowles Commission Monograph, Ch. 15. New York: John Wiley & Sons.

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.

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.

Krolzig, H.-M., and Hendry, D. F. (2003). Assessing subsample-based model selection procedures. Working paper, Economics Department, Oxford University.

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

Omtzig, P. (2002). Automatic identification and restriction of the cointegration space. Thesis chapter, Economics Department, Copenhagen University.

Owen, P. D. (2003). General-to-specific modelling using PcGets. *Journal of Economic Surveys*, forthcoming.

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

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

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