# The Properties of Automatic *Gets* Modelling

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

December 21, 2002

**Abstract**

We describe some recent developments in *PcGets*, and consider their impact on its performance across different (unknown) states of nature. We discuss the consistency of its selection procedures, and 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. We also describe how *PcGets* has been extended to assist non-experts in model formulation, handle more variables than observations and tackle non-linear models.

## Contents



*Prepared for EC$^2$, Bologna, 2002. Financial support from the U.K. Economic and Social Research Council under grant L11625015 is gratefully acknowledged. We are indebted to Dorian Owen for suggesting a correction to the degrees of freedom of the heteroscedasticity tests, and Julia Campos for comments on an earlier draft.

# 1 Introduction

Model selection theory poses great difficulties: all the statistics for selecting models and evaluating their specifications have distributions, usually interdependent, and possibly altered by every modelling decision. Fortunately, recent advances in computer automation of selection algorithms have allowed a fresh look at this old problem, both by revealing some high success rates, and by allowing operational studies of alternative strategies: see *inter alia* Hoover and Perez (1999), Hendry and Krolzig (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 some of the procedures in, and recent changes to, *PcGets*, and seek to ascertain their impact on its behaviour in sifting relevant from irrelevant variables in econometric modelling.[1] Hendry and Krolzig (2002) described the selection strategies embodied in *PcGets*, their foundation in the theory of reduction, and potential alternatives. They emphasized the distinction between the costs of inference, which are an inevitable consequence of non-zero significance levels and non-unit powers, and the costs of search, which are additional to those faced when commencing from a model that is the data generation process (DGP). Finally, they provided Monte Carlo evidence on the performance of *PcGets* in a range of experiments, including those used to calibrate its settings.

This paper provides an update on Hendry and Krolzig (2002), by considering seven recent developments. First, the consistency of the implemented model selection strategy, as embodied in *PcGets*' Liberal and Conservative strategies, is discussed. Secondly, the progress of *PcGets*, as the algorithm has been refined, is demonstrated by again re-running some of the previously published Monte Carlo experiments from Hendry and Krolzig (1999), Krolzig and Hendry (2001) and Hendry and Krolzig (2002). The associated developments are also discussed. Thirdly, we investigate the presence/absence of 'pre-test biases' and 'model selection effects', for both estimators and tests. Fourthly, we analyze the sub-sample 'significance evaluation' procedure which acts as a reliability check on the selected model. Next, an 'automatic modeller' has been implemented in *PcGets*, and we describe how it works, and why it may be able to outperform all but expert econometricians in selecting from an initial dynamic general unrestricted model (GUM). The sixth development is for the setting where there are more variables than observations, which surprisingly, is not necessarily a major problem for *PcGets*. Finally, the same logic is applied to selecting a non-linear model when the desired class is known.

The structure of the paper is as follows. Section 2 considers the consistency of the released version of *PcGets*, and compares *PcGets* to model selection based purely on the Schwarz, or Bayesian, information criterion (Schwarz, 1978). Section 3 summarizes the progress of *PcGets* on the various Monte Carlo experiments. Then section 4 investigates possible 'pre-test biases' and 'model selection effects'. Section 5 considers the sub-sample reliability assessment procedure. Section 6 describes the quick modelling facility, and section 7 comments on collinearity problems. Section 8 addresses the issue of more variables than observations and section 9 comments on selecting a non-linear model. Section 10 concludes. The appendix provides details of how *PcGets* now performs on the experiments in Lovell (1983) as re-analyzed by Hoover and Perez (1999), and in Krolzig and Hendry (2001) and Hendry and Krolzig (2002).

[1] *PcGets* is an Ox Package (see Doornik, 1999) implementing automatic general-to-specific (*Gets*) modelling for linear regression models based on the theory of reduction, as in Hendry (1995, Ch.9).

## 2 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 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, denoted *SC* (also called the Bayesian information criterion, *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 of the procedure converges on zero (so irrelevant variables are never retained). In particular, *SC* penalizes the log-likelihood by $n\log(T)/T$, whereas *HQ* uses $2n\log(\log(T))/T$, which they show is the minimum rate at which additional parameters must be penalized. 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.

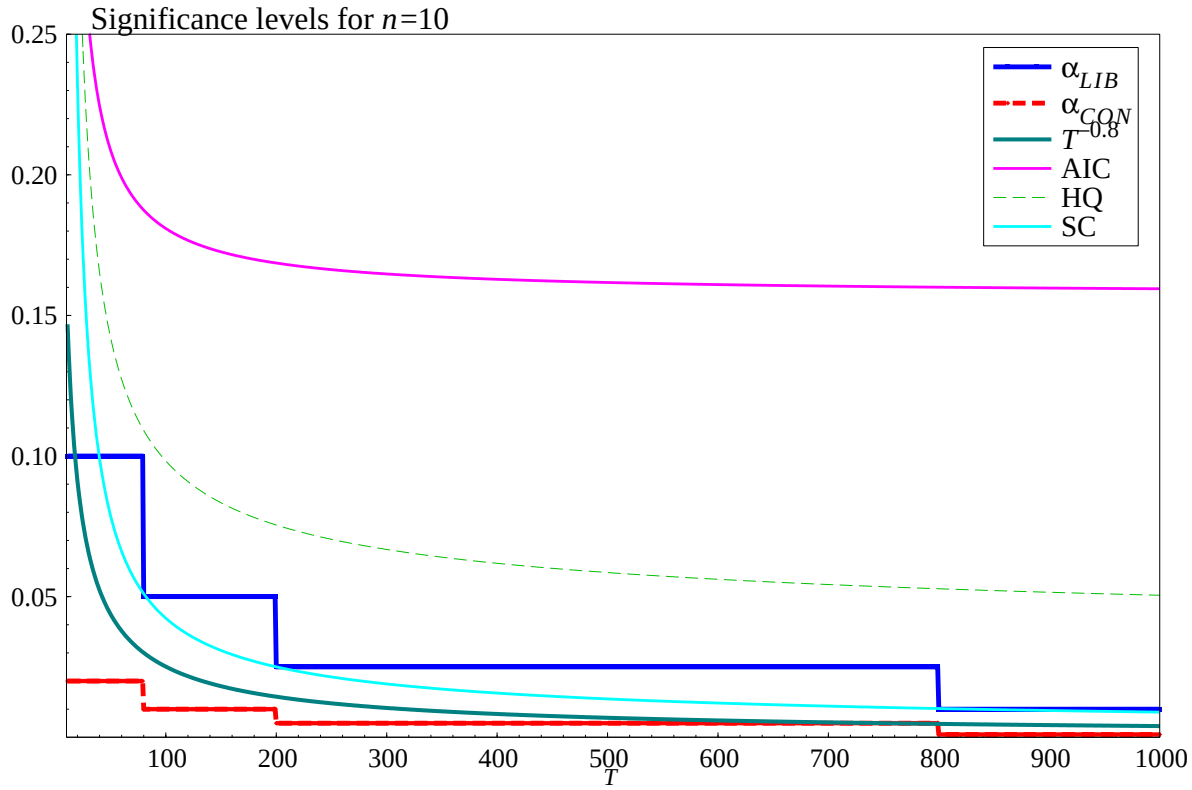


**Figure 1** Significance level comparisons across selection rules.

*PcGets* implements similar requirements to those needed for consistent selection in both its Liberal and Conservative strategies, namely the general model is assumed to be over-parameterized relative to the (local) DGP, and the nominal significance level tends to zero as the sample size increases. The Liberal strategy seeks to minimize the chances of omitting variables that matter, so uses a relatively loose significance level (with HQ as its upper and SC as its lower bound), whereas the Conservative seeks to minimize the chances of retaining variables that do not matter, and hence uses a stringent significance level. Figure 1 illustrates its rules for 10 variables (based on Hendry, 1995, Ch. 13) relative

to *AIC*, *SC* and *HQ*. As can be seen, the *PcGets* Conservative profile is much tighter than the three information criteria considered, whereas the Liberal strategy has *HQ* as its upper and *SC* as its lower bound. 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 that using $T^{-0.8}$ (also shown), but as the strategies are designed for non-expert users, it seemed preferable to base them on '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 *SC* at small samples, so the next sub-section addresses its comparison with the Schwarz criterion, viewed as *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, users ought to 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.

### 2.1 Comparisons with *BIC*

The Schwarz (1978), or Bayesian, information criterion selects from a set of $n$ candidates the model with $k$ regressors which minimizes:

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

where $c \geq 1$ and:

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

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 1 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 establish the formal link of *BIC* to significance levels.

Consider the impact of adding an extra orthogonalized regressor $z_{k+1,t}$, to the 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 (1):

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

where:

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

and:

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

Consequently:

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

so:

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

Hence, $SC_{k+1} < SC_k$ when:

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

so the additional regressor will be retained when:

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

Thus, choosing $c$ is tantamount to choosing the $\mathsf{p}$-value for the corresponding $\mathsf{t}$-test. For example, when $T=140$, with $c=1$ (the usual choice), and $k=40$, as in Hoover and Perez (1999), we have $SC_{41} < SC_{40}$ 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:

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

which is a sequence of $|\widehat{\mathsf{t}}_{(i)}|$ between 1.9 (at $k=40$) and 2.2 (at $k=1$). That 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:

$$\mathsf{P}\left(|\mathsf{t}_{(i)}| < 1.9 \ \forall i = 1,\ldots,40\right) = (1-0.0595)^{40} = 0.09. \tag{2}$$

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 140 degrees of freedom), $\mathsf{P}\left(|\mathsf{t}_{(i)}| < 2.2\right) = 0.3$. Reducing both $T$ and $k$ can worsen the chances of correct selection: for example, $T=80$, $c=1$ and $k=30$ leads to $\mathsf{P}\left(|\mathsf{t}_{(i)}| < 1.66 \ \forall i = 1,\ldots,30\right) = 0.04$. Such probabilities of correctly selecting a null model are too low to provide a useful practical basis. Two amendments have been proposed.

First, lowering the maximum size of model to be considered using 'pre-testing' as in (say) Hansen (1999). He uses $n=10$ when $T=140$ by sequentially eliminating the variable with the smallest $\mathsf{t}$-value at each stage until 30 are removed. However, that procedure entails that *BIC* actually confronts a different problem. If pre-selection did not matter, then under the null we would have:

$$\mathsf{P}\left(|\mathsf{t}_{(i)}| < 2.16 \ \forall i = 1,\ldots,10\right) = (1-0.0325)^{10} = 0.72. \tag{3}$$

But the un-eliminated variables are those selected to have the largest $\mathsf{t}$-values, so (3) overstates the performance. Conversely, (2) 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 in (2):

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

so even allowing for the initial existence of 40 variables matters considerably.

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:

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

which is a dramatic improvement over (2). 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 $(1 - 0.0078)^{10} = 0.92$).

Nevertheless, when the null is false, both steps (i.e., raising $c$ and 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: that this effect does not show up in the Hoover–Perez experiments is due to the 'population' t-values either being very large or very small.

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 DGP variables.

### 2.1.1 Comparisons in a VAR

In Brüggemann, Krolzig and Lütkepohl (2002), the *Gets* approach to the reduction of vector autoregressive models is compared to selection procedures based on information criteria. For the DGPs considered in their Monte Carlo study, the forecast comparison indicated a clear advantage for *PcGets*. The results are summarized in figure 2, which reports the relative mean squared prediction error (MSPE) of the models selected by *PcGets* and full-search procedures maximizing AIC, HQ and SC, respectively, at forecast horizons $h = 1$ and $5$. Interestingly, the forecasts produced by *PcGets* are better than the forecasts based on the true model when all non-zero coefficients of the DGP have to be estimated: in other words, the estimated DGP forecasts are affected by estimation uncertainty, whereas any model selection uncertainty is offset by the simplification gains, as might be expected from the theory in the previous section.

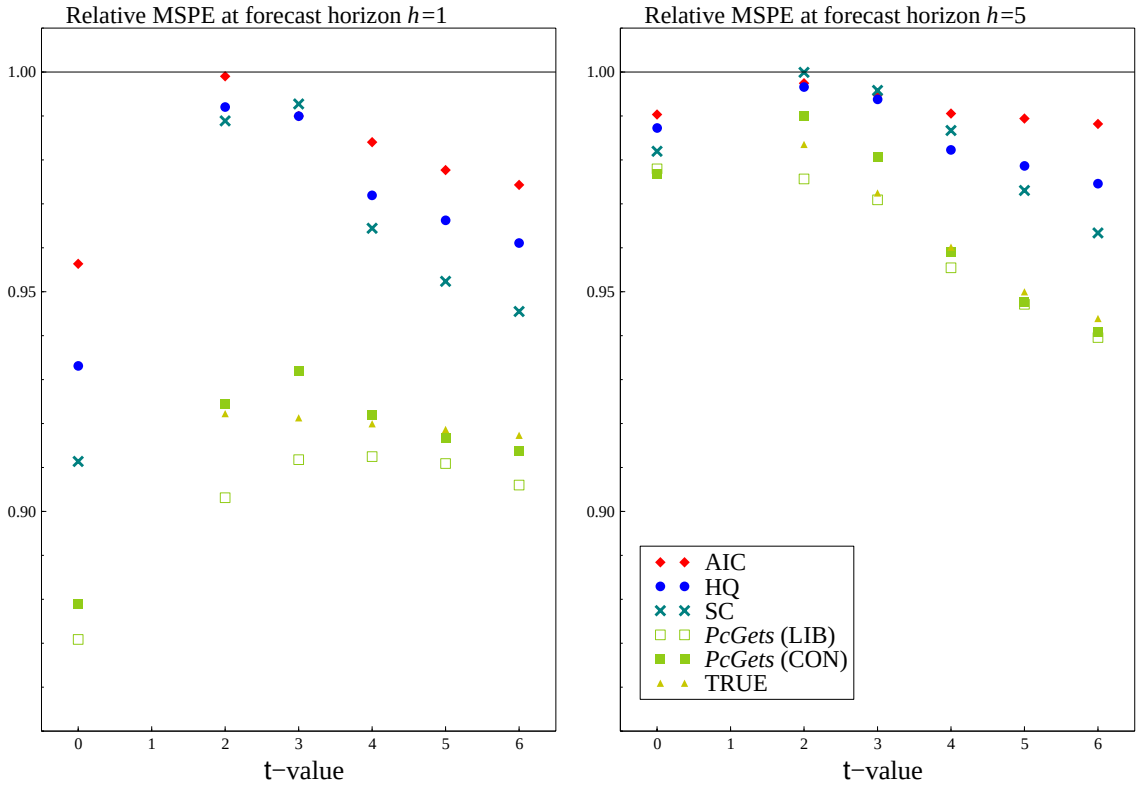


**Figure 2** Normalized and averaged MSPEs relative to the unrestricted VAR.

# 3 Progress in *PcGets*

Various new, corrected, and additional procedures have been implemented, most having only a small impact on the program's behaviour. This is unsurprising given the degree of 'error correction' manifest in the experiments used to calibrate the program (i.e., when one procedure does not perform, another does) combined with how close to the theoretical upper bound performance already is. Nevertheless, improvements are feasible in several directions. First, for settings not previously envisaged, such as a model with (say) forty lags of one variable, and few lags on others. When one important effect is hidden in a morass of irrelevance, the pre-search block tests need not be appropriate, so we consider using the outcome of the maximum t-test as a check (sub-section 3.1). Secondly, the calibration of the mis-specification heteroscedasticity tests was poor in Hendry and Krolzig (2002), but this transpires to be a problem with the degrees of freedom assumed for the reference distribution (sub-section 3.2). Thirdly, a number of small changes have been implemented, including one to the determination of the lag order, using a combined top-down/bottom-up approach, complemented by an automatic Lagrange-multiplier (LM) test for omitted regressors. We also investigated exploiting the information in the ordered t-statistics to locate a cut-off between included and excluded variables, but while suitable for orthogonal problems, multi-path search remains necessary in general: section 7 briefly addresses the collinearity issue. 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 the *PcGets* calibration experiments in Hendry and Krolzig, 2002).

## 3.1 Max t-tests

When only one of a large set $n$ of candidate variables matters, then on average, a block test $F_{T-n}^{n}$ will have low power to detect it compared to a focused t-test. A crude approximation relating these statistics,

**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 |

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, but ignoring sampling variation:

$$\mathsf{F}^n_{T-n} \simeq \frac{1}{n}\sum_{i=1}^{n-1} 1 + \frac{1}{n}\mathsf{t}^2_{(n)} = 1 + \frac{1}{n}\left(\mathsf{t}^2_{(n)} - 1\right), \tag{4}$$

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 $\alpha$, then a $\max\{|t|\}$ criterion with the correct size would use the approximate nominal significance level (see e.g., Savin, 1984):

$$\delta^\alpha_n = 1 - (1-\alpha)^{1/n}. \tag{5}$$

For example, for $n = 10$ when $\alpha = 0.05$ so $\mathsf{P}\big(\mathsf{F}^{10}_{100} > 1.93|\mathsf{H}_0\big) = 0.05$, then from (4), a significant outcome due to only $\mathsf{t}^2_{(10)}$ requires its value to be about 10.3, whereas from (5):

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

which entails $\mathsf{t}^2 > 8.1$, and so is somewhat smaller. Nevertheless, one relevant variable can easily hide in a set where the overall outcome is insignificant: and note the potential for conflicting inference—*PcGets* judges the variable as irrelevant by the F test or a t-test based on $\delta^\alpha_n$, whereas a later investigator using a one-off t-test at significance level $\alpha$ would include it. Thus, a compromise between size and power more favourable to the latter when the initial specification is highly over-parameterized, but one or more of the variables matters, is to consider the $\max\{|t|\}$ statistic, but at a less stringent level than $\delta^\alpha_n$, say twice the value from (5).

To investigate the quality of the approximation in (4), we consider 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 $M = 100000$ replications, we calculate the maximum $\max\{|\mathsf{t}_1|, \ldots, |\mathsf{t}_n|\}$ of the $n$ random variables, and compare the t-prob of its $1-\alpha$ quantiles to the prediction of the $\delta^\alpha_n$ rule. Figure 3 plots $\delta^\alpha_n$ for $\alpha = 0.01$ and $0.05$ and compares it to the $0.95$ and $0.99$ quantiles of associated t-probs The results demonstrate the quality of the approximation.

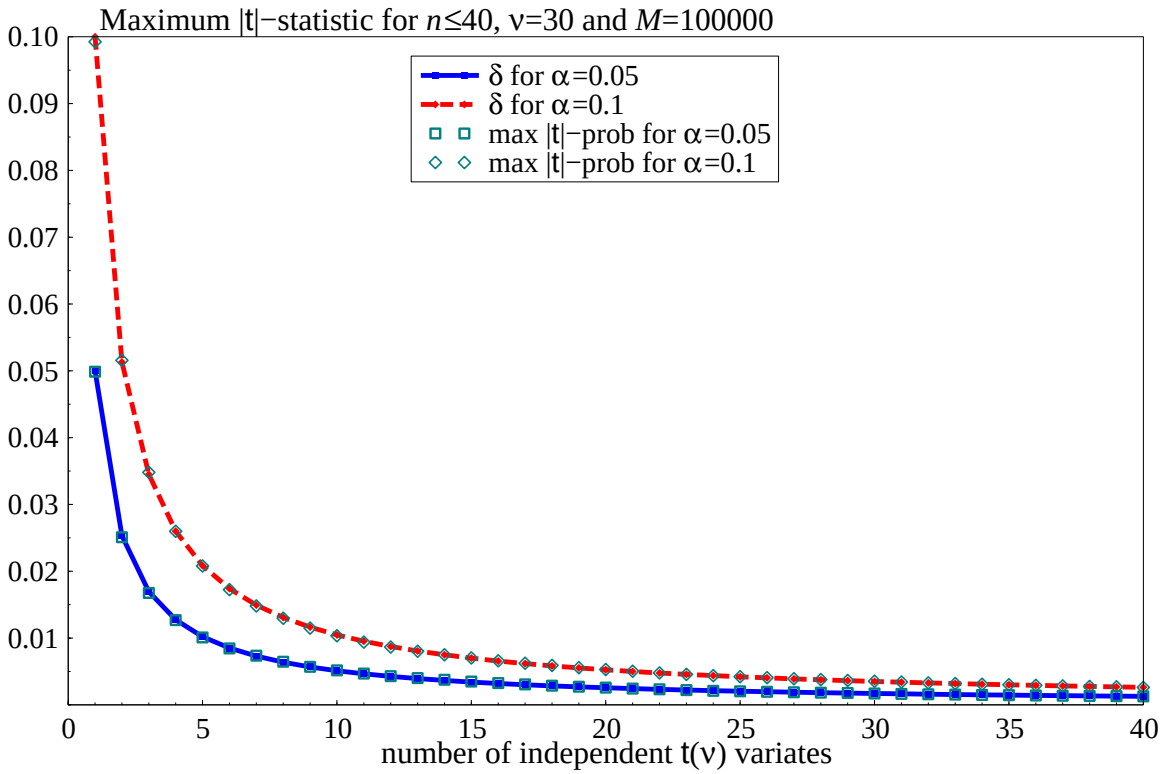


**Figure 3** $\delta_n^\alpha$ and $\max|\mathsf{t}|$ of $n$ IID $\mathsf{t}(\nu)$ random variates.

### 3.2 Recalibrating the heteroscedasticity tests

In Krolzig and Hendry (2001), we 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 we therefore cautioned against their use. A reviewer of Hendry and Krolzig (2001) (Dorian Owen) suggested that the degrees of freedom were inappropriate by using a correction like that in Lagrange-multiplier autocorrelation tests (see e.g., Godfrey, 1978, 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. This indeed produces much closer matches with their anticipated distributions as tables 2 and 3 show for the ARCH and heteroscedasticity tests applied to the DGP.

**Table 2** ARCH test DGP outcomes.

| Nominal | 10% | 7.5% | 5% | 2.5% | 1% |
|---|---|---|---|---|---|
| HP2 | 0.075 | 0.058 | 0.039 | 0.018 | 0.006 |
| HP2* | 0.069 | 0.055 | 0.037 | 0.020 | 0.008 |
| HP7 | 0.068 | 0.054 | 0.036 | 0.015 | 0.005 |
| JEDC | 0.104 | 0.076 | 0.053 | 0.025 | 0.009 |
| S2 – S4 | 0.064 | 0.049 | 0.035 | 0.021 | 0.008 |
| S2* – S4* | 0.095 | 0.066 | 0.046 | 0.028 | 0.015 |

The ARCH test remains under-sized in these experiments at most quantiles, whereas the heteroscedasticity test is close to its nominal significance level in most cases. Overall, there is a marked

**Table 3** Heteroscedasticity test DGP outcomes.

| Nominal | 10% | 7.5% | 5% | 2.5% | 1% |
|---|---|---|---|---|---|
| HP2* | 0.082 | 0.061 | 0.039 | 0.020 | 0.010 |
| HP2 | 0.075 | 0.053 | 0.034 | 0.019 | 0.008 |
| HP7 | 0.084 | 0.066 | 0.048 | 0.030 | 0.016 |
| JEDC | 0.095 | 0.074 | 0.054 | 0.031 | 0.016 |
| S2 – S4 | 0.097 | 0.078 | 0.055 | 0.032 | 0.013 |
| S2* – S4* | 0.092 | 0.072 | 0.052 | 0.029 | 0.016 |

improvement compared to the outcomes reported in Krolzig and Hendry (2001).

Next, we consider the improvements in the simulation behaviour of *PcGets*.

### 3.3 Overview of progress to date

As Hendry and Krolzig (2002) provide a relatively recent review of *PcGets* (as of June 2001), we record the detailed outcomes in the Appendix (section 11), and summarize the findings here. Since the study of automatic selection procedures began, progress has been substantial. First, we consider control over 'size', such that the actual null rejection frequencies are close to the nominal level set by the user 'independently' of the problem investigated. Figure 4 plots the ratio of actual to nominal size, across the various studies of the Hoover–Perez experiments at 5% and 1% nominal level, and shows that stabilization has occurred as we have learned more about how *PcGets* functions, and added new features to its search procedures, such as those noted above (the latest estimates incorporate the sub-sample reliability weightings, and are slightly under-sized—despite their being between 35 and 40 irrelevant regressors).

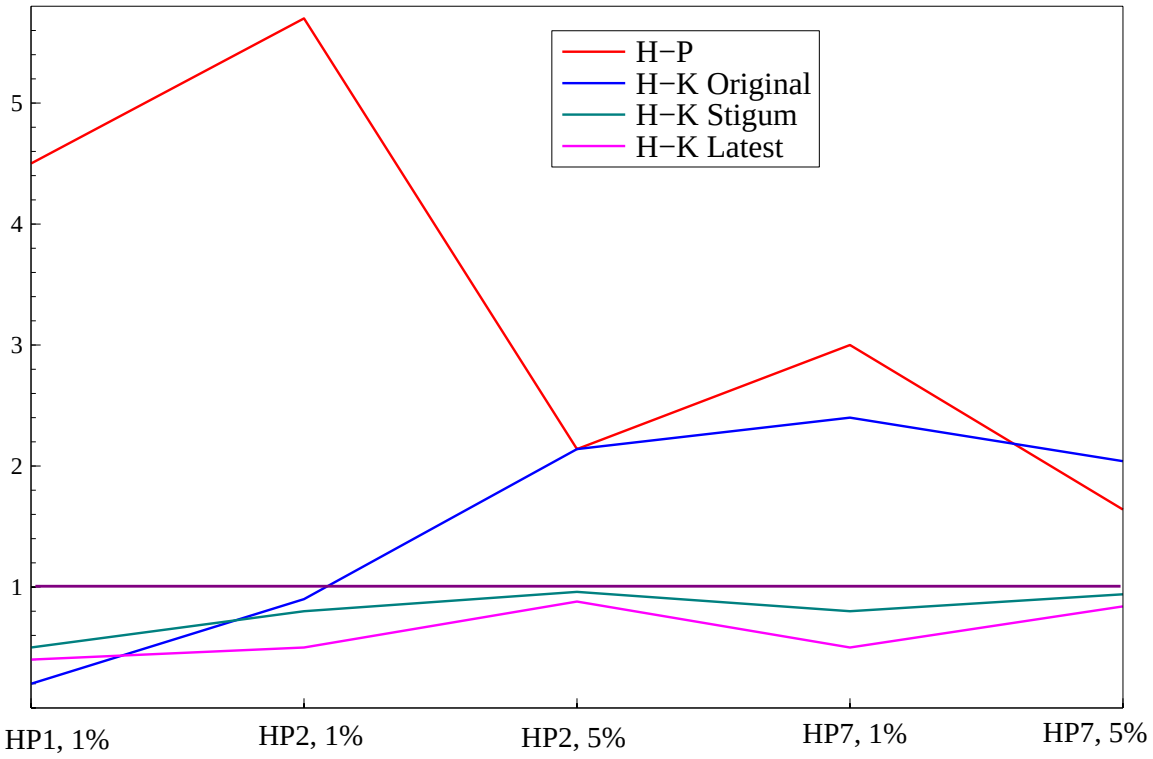


**Figure 4** Ratio of actual to nominal size.

Secondly, we consider the appropriate calibration of the two basic strategies. Figure 5 graphically illustrates four main aspects of the outcomes across all the Monte Carlo experiments to date for both Conservative and Liberal strategies. Panel a concerns 'unbiased' fit, in the sense that the final estimate of the equation standard error ($\widehat{\sigma}$) is close to the true value. The Liberal strategy has a slight downward bias (less than 5%) whereas the Conservative is upward biased by a similar amount. Such behaviour is easily explained: the latter eliminates variables which matter and the former retains some which only do so by chance. However, at no stage was selection based on fit *per se*, although a minimal congruent encompassing model will necessarily have the best fit at the significance level set.

Panel b shows sizes for the strategies relative to their intended significance levels of 5% and 1%, both with and without sub-sample reliability weightings: the latter are close to their targets.

Panel c considers the impact of the sub-sample reliability weightings on the resulting power, and shows that there is only a small effect, even at quite low powers where it should have most impact. The Conservative strategy naturally has no higher power than the Liberal, and shows that the cost of avoiding spurious variables can be high in terms of missing what matters.

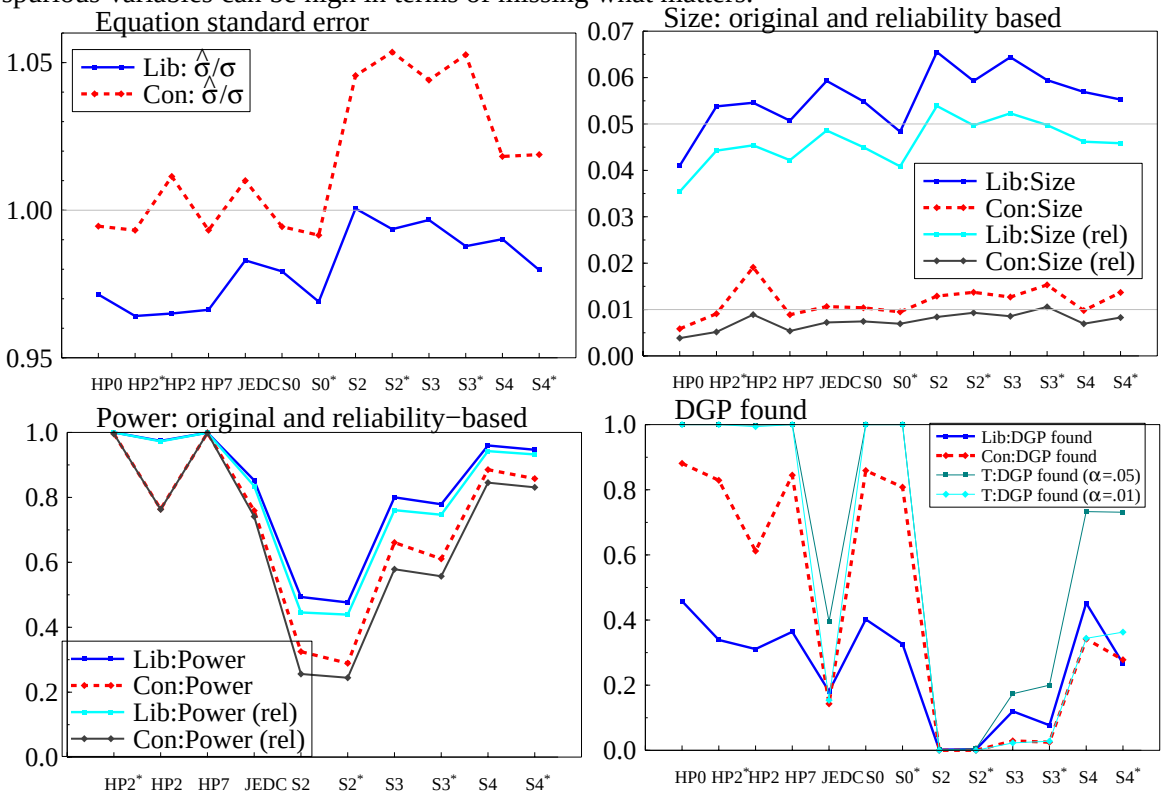


**Figure 5** Overview of accuracy, size, power and success.

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. The out-performance of commencing from the DGP in the Hoover–Perez experiments is owing to the high degree of over-parameterization but even so, the Conservative strategy does a respectable job. When population t-values are 2 or 3, the Liberal strategy does well, and sometimes outperforms commencing from the DGP with a 1% significance level. 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.

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

# 4 'Pre-test' and 'selection' biases

To investigate the impact of selection on coefficient estimates and standard errors, we recorded these outcomes in the Hendry and Krolzig (2002) experiments, with the results shown in table 4. As expected, conditional on being retained, the coefficient estimates are 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. Unconditionally, coefficient estimates are downward biased. More importantly, the estimated standard errors are not biased on either strategy, although the sampling standard deviations are. As noted earlier, the equation standard error is, if anything, upward biased, so any accusation that *PcGets* 'overfits' is clearly false. Finally, 'pre-test' effects are not changed by search *per se*, as the coefficient biases are closely similar when commencing from the DGP and the GUM.

**Table 4** Coefficient estimates and SEs.

| | 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 | | | | | | | |
| $\sigma$ | 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% | |

### 4.1 Selection effects on the two heteroscedasticty tests

Another feature of interest is the impact of model selection on the outcomes of test statistics. This is shown in tables 5 and 6 for the two tests in section 3.2. 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 $0.5\%$.

**Table 5** ARCH test selected model outcomes.

| Nominal | 10% | 7.5% | 5% | 2.5% | 1% |
|---|---|---|---|---|---|
| HP2 | 0.067 | 0.050 | 0.032 | 0.014 | 0.001 |
| HP2* | 0.058 | 0.046 | 0.032 | 0.016 | 0.002 |
| HP7 | 0.054 | 0.042 | 0.024 | 0.009 | 0.002 |
| JEDC | 0.103 | 0.076 | 0.043 | 0.011 | 0.001 |
| S2 | 0.070 | 0.048 | 0.033 | 0.016 | 0.001 |
| S3 | 0.072 | 0.058 | 0.037 | 0.016 | 0.000 |
| S4 | 0.066 | 0.051 | 0.029 | 0.016 | 0.007 |
| S2* | 0.067 | 0.049 | 0.027 | 0.011 | 0.001 |
| S3* | 0.082 | 0.058 | 0.044 | 0.024 | 0.001 |
| S4* | 0.088 | 0.057 | 0.040 | 0.019 | 0.003 |

For the Hoover–Perez DGPs, the heteroscedasticity test statistics were all insignificant at $10\%$. While the regressors in the JEDC and S experiments are generated by a Gaussian white-noise process, the regressors in the HP experiments are heteroscedastic.

**Table 6** Heteroscedasticity test selected model outcomes.

| Nominal | 10% | 7.5% | 5% | 2.5% | 1% |
|---|---|---|---|---|---|
| JEDC | 0.109 | 0.084 | 0.056 | 0.027 | 0.004 |
| S0 | 0.011 | 0.009 | 0.006 | 0.003 | 0.001 |
| S2 | 0.108 | 0.083 | 0.057 | 0.028 | 0.004 |
| S3 | 0.116 | 0.090 | 0.061 | 0.027 | 0.004 |
| S4 | 0.107 | 0.082 | 0.056 | 0.028 | 0.007 |
| S0* | 0.013 | 0.010 | 0.006 | 0.003 | 0.000 |
| S2* | 0.098 | 0.077 | 0.052 | 0.026 | 0.002 |
| S3* | 0.111 | 0.087 | 0.057 | 0.026 | 0.003 |
| S4* | 0.104 | 0.080 | 0.057 | 0.026 | 0.003 |

As can be seen in comparison with tables 2 and 3 above, 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 tables do not represent a 'distortion' of the sampling properties: the key decision is that taken at the level of the general model, and conditional on not rejecting there, no change should occur in that decision. Tables 7 and 8 confirm that result: in both cases, the tests have their anticipated operating characteristics.

## 5 Sub-sample reliability assessment

After selection, the relevance of variables in the final model is explored by post-selection reliability checks to ascertain whether 'significance' is substantive or adventitious. Post-selection evaluation is an attempt to mimic the role in an automatic procedure of recursive estimation, aiming to evaluate whether apparently significant effects are substantive or chance. It is not a check on constancy, which has already been tested for the GUM, and checked by diagnostics at each potential reduction.

**Table 7** ARCH test general model outcomes.

| Nominal | 10% | 7.5% | 5% | 2.5% | 1% |
|---|---|---|---|---|---|
| HP0 | 0.146 | 0.118 | 0.095 | 0.053 | 0.033 |
| HP2 | 0.114 | 0.093 | 0.066 | 0.038 | 0.022 |
| HP2* | 0.106 | 0.085 | 0.060 | 0.039 | 0.024 |
| HP7 | 0.123 | 0.102 | 0.074 | 0.043 | 0.024 |
| JEDC | 0.074 | 0.063 | 0.041 | 0.022 | 0.011 |
| S0 – S4 | 0.074 | 0.060 | 0.039 | 0.018 | 0.009 |
| S0* – S4* | 0.065 | 0.051 | 0.034 | 0.027 | 0.008 |

**Table 8** Heteroscedasticity test general model outcomes.

| Nominal | 10% | 7.5% | 5% | 2.5% | 1% |
|---|---|---|---|---|---|
| JEDC | 0.094 | 0.072 | 0.047 | 0.025 | 0.012 |
| S0 | 0.098 | 0.076 | 0.055 | 0.029 | 0.014 |
| S2 | 0.098 | 0.077 | 0.054 | 0.028 | 0.012 |
| S3 | 0.100 | 0.078 | 0.055 | 0.028 | 0.013 |
| S4 | 0.099 | 0.077 | 0.055 | 0.029 | 0.014 |
| S0* | 0.090 | 0.068 | 0.047 | 0.022 | 0.010 |
| S2* | 0.089 | 0.068 | 0.047 | 0.024 | 0.010 |
| S3* | 0.090 | 0.068 | 0.046 | 0.024 | 0.010 |
| S4* | 0.090 | 0.067 | 0.045 | 0.023 | 0.010 |

Under the null hypothesis $\mathsf{H}_0$, using a 2-sided test, a t-value will exceed (in absolute value) a critical value $c_\alpha$ on $\alpha\%$ of the occasions, where $\alpha$ is the significance level, so:

$$\mathsf{P}\left(-c_\alpha \leq \mathsf{t} \leq c_\alpha \mid \mathsf{H}_0\right) = \alpha.$$

However, after selecting a model, the retained variables will have significant t-values by construction.[2] The selected set thus comprises (on average) $\alpha\%$ of the initial set—significant by chance—and the remainder—significant by having non-central t-distributions. The issue is whether conditional on observing full-sample significance, there is a division of the sample into sub-samples that would help discriminate between these, exploiting the fact that non-central t-values diverge, whereas central t-values are only significant by a chance value falling outside the range $[-c_\alpha, c_\alpha]$ at the end of the sample.

Our proposed filter between variables that really matter (non-central ts) and those that are adventitiously significant (central ts that happen to take large end-of-period values) is to check sub-sample reliability. The idea is that the central t-tests will be low in at least one of the two sub-periods, so revealing the actual irrelevance of the associated variable. Because the sample sizes are smaller, less stringent critical values are used. Hoover and Perez (1999) find evidence that the power-size trade-off as a function of the sample split is 'flat' in the neighbourhood of 75–25 splits (so 50% of observations are in common), hence *PcGets* centers on that.

It is clear from all the Monte Carlo studies we have conducted that the reliability check reduces the size, and has helped stabilize performance over different states of nature. Nevertheless, that by itself does not resolve the key issue of whether an equivalent size reduction achieved by lowering the initial

[2] We neglect the small percentage of the time where variables enter insignificantly, but their elimination would induce a significant diagnostic test value.

significance level of every test would result in higher or lower power, and if so, how that changes across different DGPs. The size-power trade-off is highly non-linear in both the significance level and the non-centrality parameters of the variables, and any analysis must be conditional on having retained each associated regressor at its observed t-value.

We have undertaken extensive analytic investigations of the problem, but have few clear results at this stage. However, we have shown that a 50–50 split (where the sub-samples are independent draws) yields no benefit, and is equivalent in terms of mis-classifying relevant and irrelevant variables to the same reduction in the significance level imposed in the full sample. Conversely, since the simulation evidence in Hoover and Perez (1999) suggested that a 50–50 split was far from optimal, overlapping samples might yet be proved to deliver genuine gains.

We also investigated the sub-sample properties of a single t test when the analysis is conditioned on its significance in the full sample. The Monte Carlo study consists of 5,000,000 replications of the experiment with two $\mathsf{t}(\nu, \psi)$ distributed random variables with $\nu = 50$ degrees of freedom and a non-centrality $\psi \in \{0, 2, 3, 4\}$. The full-sample $|\mathsf{t}|$-value is given by $|\mathsf{t}| = \frac{1}{\sqrt{2}} |\mathsf{t}_1 + \mathsf{t}_2|$.

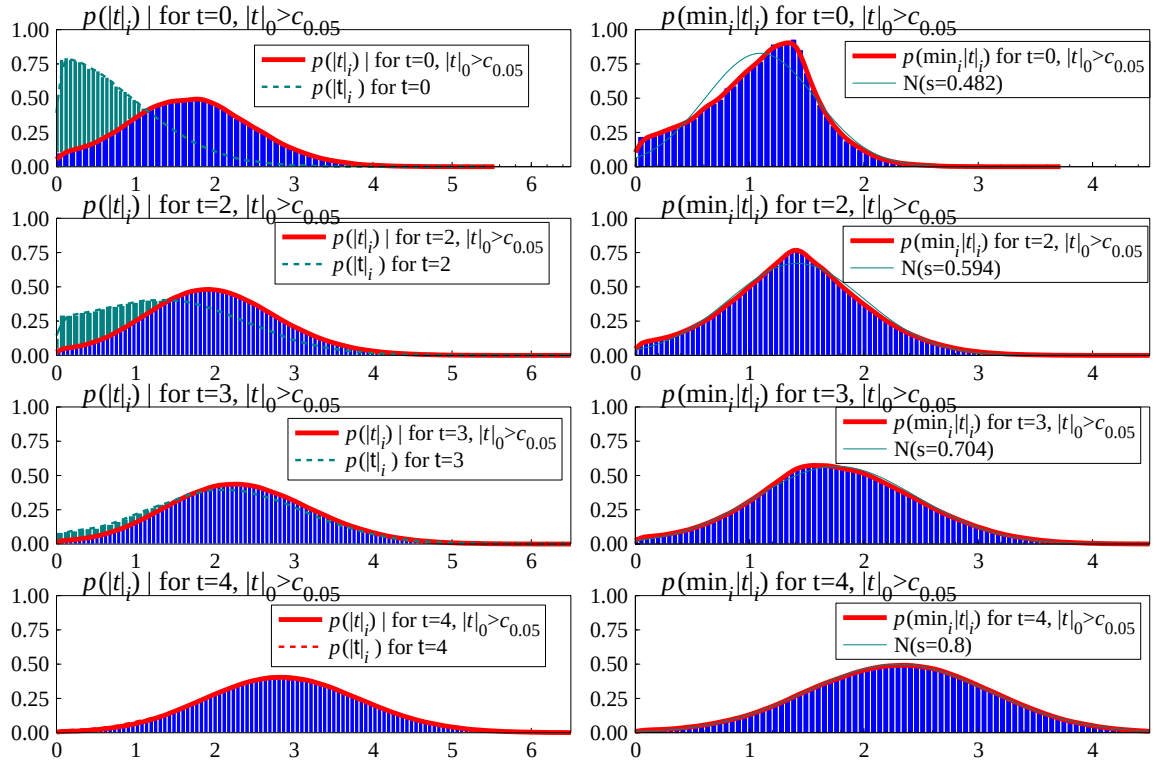


**Figure 6** The density of $|\mathsf{t}|_i$ and $\min_i\{|\mathsf{t}|_i\}$— conditional on significance in the full sample.

Figure 6 plots the conditional density of $|\mathsf{t}|_i$ and $\min_i\{|\mathsf{t}|_i\}$— in non-overlapping subsamples conditional on significance in the full sample. It is evident that if a regressor is significant in the full sample, the distribution of the subsample $|\mathsf{t}|$-values of a variable that matters is hardly distinguishable from that of a nuisance variable. Information from overlapping subsamples is required for the reliability statistic. In the split-sample analysis of *PcGets*, the size of the subsample is $0.75T$.

It is important to distinguish the reliability assessment of a model (which has been selected based on the full-sample information) from selection rules that are formulated in terms of sub-sample evidence. Hoover and Perez (1999) proposed selecting only variables that are significant in two (over-lapping) subsamples. We now provide some Monte Carlo evidence indicating that the latter procedure is dominated

by the former.[3]

Let $\{\mathtt{t}_1, \mathtt{t}_2\}$ be $\mathtt{t}(\nu)$ distributed random variables. Figure 7 compares the distribution of the full-sample $|\mathtt{t}| = \frac{1}{\sqrt{2}}|\mathtt{t}_1 + \mathtt{t}_2|$ and the minimum of the $|\mathtt{t}|$-values, $\min\{|\mathtt{t}|_1, |\mathtt{t}|_2\}$, in the two non-overlapping sub-samples of size $\nu_1 = \nu_2 = 50$.

**Table 9** Power function: sub-sample $\min\{|\mathtt{t}|_1, |\mathtt{t}|_2\}$ or full-sample t test.

| t-value | full-sample | | | subsample $(0.5T)$ | | | subsample $(0.75T)$ | | | subsample $(0.85T)$ | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 10% | 5% | 1% | 10% | 5% | 1% | 10% | 5% | 1% | 10% | 5% | 1% |
| 2 | 0.624 | 0.498 | 0.258 | 0.443 | 0.333 | 0.162 | 0.562 | 0.438 | 0.213 | 0.588 | 0.461 | 0.228 |
| 3 | 0.902 | 0.837 | 0.633 | 0.750 | 0.661 | 0.454 | 0.852 | 0.770 | 0.541 | 0.874 | 0.796 | 0.574 |
| 4 | 0.988 | 0.975 | 0.908 | 0.928 | 0.889 | 0.765 | 0.973 | 0.947 | 0.841 | 0.980 | 0.960 | 0.870 |
| 5 | 0.999 | 0.998 | 0.989 | 0.985 | 0.974 | 0.933 | 0.997 | 0.993 | 0.969 | 0.998 | 0.996 | 0.979 |
| 6 | 1.000 | 1.000 | 0.999 | 0.998 | 0.996 | 0.987 | 1.000 | 1.000 | 0.997 | 1.000 | 1.000 | 0.998 |

Table 9 reports the resulting power functions. It is worth noting that analyzing overlapping sub-samples can retrieve parts of the power loss. This is illustrated in figure 8 for subsample sizes of $0.75T$ and $0.85T$.

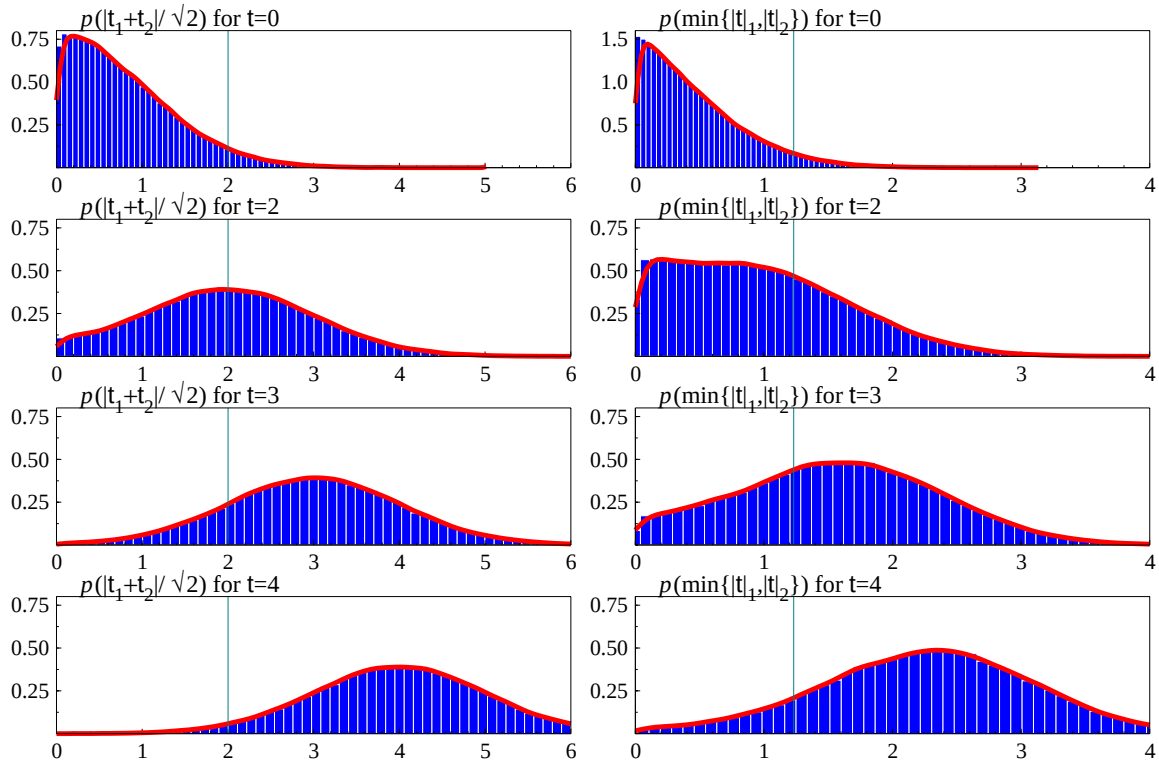


**Figure 7** Density of the full-sample $|\mathtt{t}|$ and $\min\{|\mathtt{t}|_1, |\mathtt{t}|_2\}$ for $\nu_1 = \nu_2 = 50$.

[3]Lynch and Vital-Ahuja (1998) analyzed the related problem whether the use of subsample evidence can mitigate the potential impact of data snooping on the distribution of test statistics. Comparing subsample and entire sample $R^2$ tests, Lynch finds that the full-sample test has a less distorted size and more power than the multisample test.

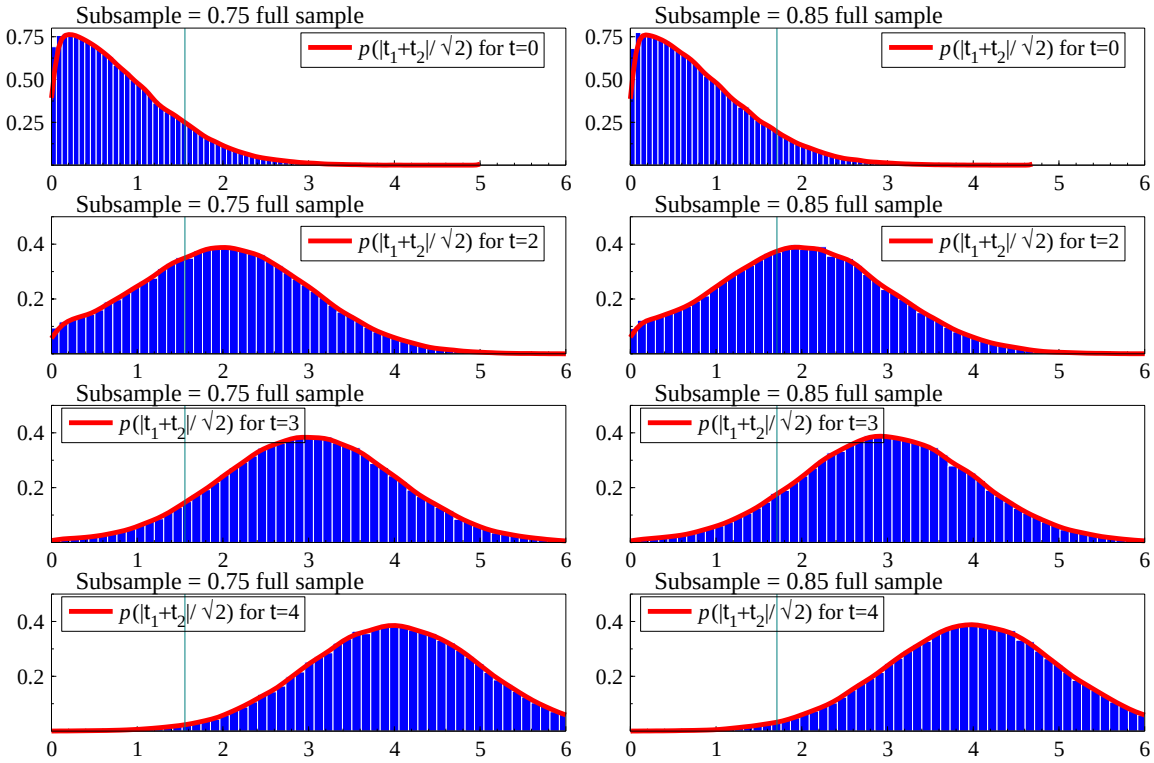


**Figure 8** Density of $\min\{|\mathsf{t}|_1, |\mathsf{t}|_2\}$ in overlapping subsamples.

## 6 Quick modeller

The latest version of the program offers a quick modelling option. The user only needs to specify the regressand and the list of basic regressors, after which *PcGets* offers a menu for fitting a static, dynamic or cross-section model. The first takes the equation as the basic list; the second selects the maximum lag length at a convex combination of (i) one more than the data frequency $f$; (ii) 0.4 time the number of observations per regressor; and (iii) the log of the number of observations $T$:

$$p^* = \left[\max\left\{\left(\min\left\{1.5+f, \frac{0.4T}{1+n}-1\right\}\right)^{\frac{1}{2}}\left(\frac{T^{\frac{3}{4}}}{1+n}-1\right)^{\frac{1}{4}}(\log T)^{\frac{1}{4}},\ 0\right\}\right],$$

(or it can be set by the user); and the third abstracts from time-related tests. Contemporaneous variables can be included or excluded, and outlier corrections implemented if desired. The Liberal strategy with sub-sample analysis is the default, after which *PcGets* creates the GUM and selects a model.

The main difference from standard 'expert usage' is that the program chooses the lag length in dynamic models. We assume the user has thought carefully about the specification—indeed, she will have more time to do so given other tasks are much less onerous—including the relevant variables and appropriate functional forms. Subject to that, its performance should be similar to more advanced usage. For example, on the DHSY data set commencing from just the list of $c$, $y$, $p$ it selects the model reported by Davidson, Hendry, Srba and Yeo (1978). Thus, while users expert in dynamic empirical modelling, willing to explore the many possible reduction paths, and with specific knowledge about the problem under analysis may well 'beat' the program, the authors' experience to date is that *PcGets* provides baseline models that are highly competitive. The main *caveat* is an expert's ability to transform the variables to near-orthogonal, interpretable representations, so we briefly reconsider the issue of collinearity.

# 7 Collinearity

Perfect collinearity denotes an exact linear dependence between variables; perfect orthogonality denotes no linear dependencies; and any state in between 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 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, even the 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 of 0.99 can arise in systems with unit roots and drift, but there is little difficulty determining the relevance of variables when the DGP is:

$$y_t = \gamma + y_{t-1} + \epsilon_t \text{ with } \epsilon_t \sim \mathsf{IN}\left[0, \sigma_\epsilon^2\right], \tag{6}$$

and the fitted model is (say):

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

and $z_t$ is generated as a random walk with drift, but independently of (6).

Conversely, in a bivariate normal:

$$\begin{pmatrix} x_t \\ z_t \end{pmatrix} \sim \mathsf{IN}_2\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right], \tag{7}$$

with a conditional model as the DGP:

$$y_t = \beta x_t + \gamma z_t + \epsilon_t \tag{8}$$

when $\rho = 0.99$ there is almost no hope of determining which variables matter in (8).

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 (8) given (7) 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 $t^2_{(i)}$ 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:

$$\begin{array}{rclcrcl} y_t & = & \sum_{k=1}^{5} \beta_{k,0} x_{k,t} + u_t, & & u_t & \sim & \mathsf{IN}\left[0, \sigma_u\right], \\ x_t & = & (\alpha \mathbf{I}_{10}) x_{t-1} + v_t, & & v_t & \sim & \mathsf{IN}_{10}\left[\mathbf{0}, \frac{\sigma_v^2}{1-\alpha}\mathbf{I}_{10}\right] \text{ for } t = 1, \ldots, T. \end{array} \tag{9}$$

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:

$$\mathtt{t}_k = \beta_k \sqrt{T \frac{\sigma_x}{\sigma_u}} = \beta_k \sqrt{T \frac{\sigma_v}{(1-\alpha^2)\sigma_u}}$$

The DGP is designed to ensure invariant population t-values with increasing $\alpha$. 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 $x_{6,t}, \ldots, x_{10,t}$ and the first lags of all regressors:

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

In an alternative experiment, we consider the orthogonal representation of (10):

$$y_t = \pi_{0,0} + \pi_{0,1} y_{t-1} + \sum_{k=1}^{10} \pi_k x_{k,t} + \sum_{k=1}^{10} \gamma_k \left(\alpha x_{k,t} - x_{k,t-1}\right) + u_t, \quad u_t \sim \mathsf{IN}\left[0, \sigma^2\right]. \tag{11}$$

In (10) as in (11), 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 (11), 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 (10) with $T = 100$, (ii) GUM (11) with $T = 100$, and (iii) GUM (11) 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 $\alpha$-invariant if the sample size is adjusted.

## 8 Selection with too many regressors

Consider two groups of variables relevant to determining a variable of interest $y$, denoted $\mathbf{x}_{i,t}$, for $i = 1, 2$, of dimensions $n_i << T$ respectively where $n = n_1 + n_2 > T$, but any one ($n_1$ or $n_2$) is sufficiently smaller than $T$ to be estimable. The analysis is easily generalized for more groups, although the computational burden rises in a combinatorial fashion. Further partition each of $\mathbf{x}_{1,t}$ and $\mathbf{x}_{2,t}$ into two halves, producing four groups. Now select (say) the first halves of $\mathbf{x}_{1,t}$ and $\mathbf{x}_{2,t}$ as the first GUM, then the second halves (assuming the ordering is arbitrary), then the cross-pairing. Cumulate all the resulting terminal models from each of those searches as the next GUM. There are ${}_2C_4 = 6$ combinations [(1,2) (3,4), (1,3) (2,4), (1,4), (2,3)] to be investigated, but the procedure is easily automated. Many of the elements in each set need not have an effect, but we assume components of each are relevant. We assume all sub-models are congruent against own information, but if non-congruent, HAC standard errors could be used. We now explain the procedure for two subsets.

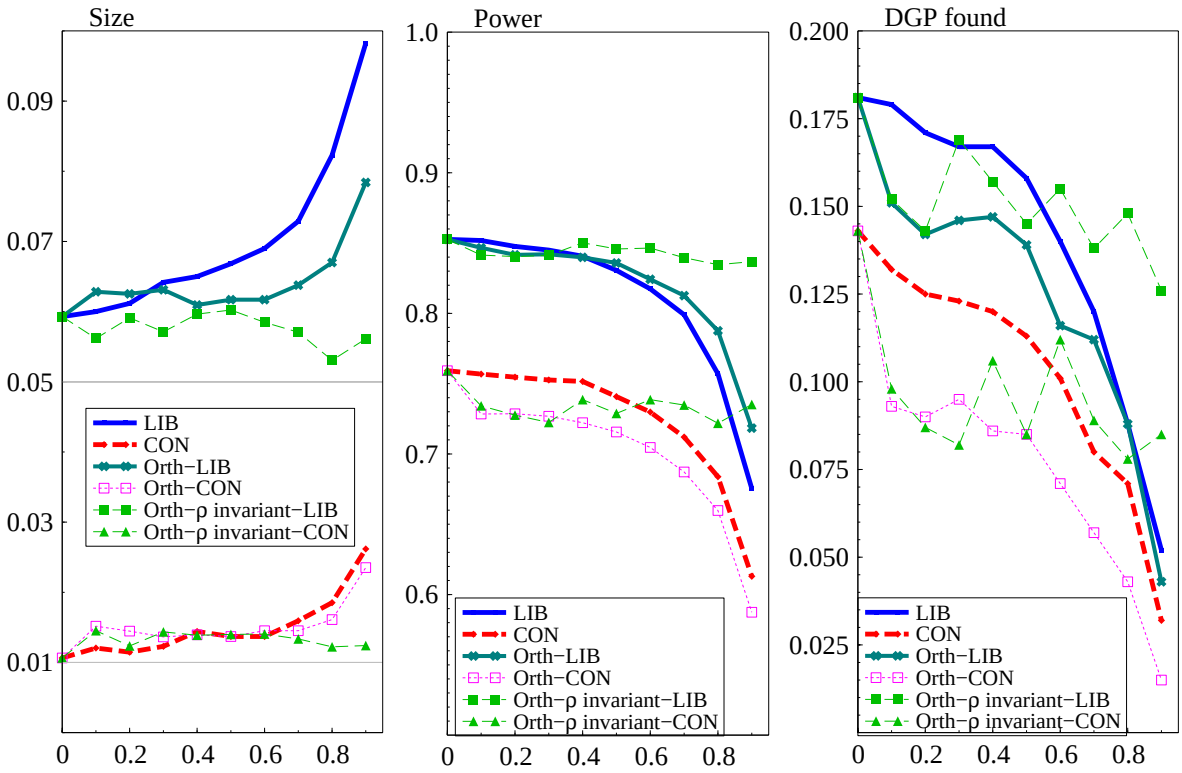


**Figure 9** Selection properties of *PcGets* for varying $\alpha$.

For $t = 1, \ldots, T$, let the DGP be:

$$y_t = \sum_{i=1}^{2} \boldsymbol{\beta}_i' \mathbf{x}_{i,t} + \epsilon_t \text{ where } \epsilon_t \sim \mathsf{IN}\left[0, \sigma_\epsilon^2\right], \tag{12}$$

where the $\boldsymbol{\beta}_i$ contain many zeros, such that the remaining non-zero parameters number $k << T$. To refer unambiguously to the signs of the covariances between variables, we take all the parameters in $\{\boldsymbol{\beta}_i\}$ as positive without loss of generality. The central case is where the groups are equal sized, so two 'general models' are considered of the form ($\underset{c}{\sim}$ denotes 'claimed to be distributed as'):

$$y_t = \boldsymbol{\gamma}_j' \mathbf{x}_{j,t} + u_{j,t} \text{ where } u_{j,t} \underset{c}{\sim} \mathsf{IN}\left[0, \sigma_{u_j}^2\right], \tag{13}$$

Then as both $\boldsymbol{\beta}_i \neq \mathbf{0}$, the selected model from each of (13) will not coincide with that selected from (12) when the latter is estimable. Nevertheless, we assume that all the models are congruent against their own information sets, perhaps by design. If (13) cannot be estimated, sub-divide further; however, two sets explains the logic.

First select the best model from:

$$y_t = \boldsymbol{\gamma}_1' \mathbf{x}_{1,t} + u_{1,t} \tag{14}$$

to get the first terminal model:

$$y_t = \boldsymbol{\lambda}_1' \mathbf{x}_{1,t}^* + v_{1,t} \text{ where } v_{1,t} \underset{c}{\sim} \mathsf{IN}\left[0, \sigma_{v_1}^2\right], \tag{15}$$

where $\mathbf{x}_{1,t}^*$ denotes the retained components of $\mathbf{x}_{1,t}$ such that all elements of $\boldsymbol{\lambda}_1$ are non-zero.

Similarly for $\mathbf{x}_{2,t}$, commence from:

$$y_t = \boldsymbol{\gamma}_2' \mathbf{x}_{2,t} + u_{2,t} \tag{16}$$

to get a second terminal model:

$$y_t = \boldsymbol{\lambda}_2' \mathbf{x}_{2,t}^* + v_{2,t} \text{ where } e_{1,t} \underset{c}{\sim} \mathsf{IN}\left[0, \sigma_{v_2}^2\right]. \tag{17}$$

Now, re-start the selection from:

$$y_t = \boldsymbol{\theta}_1' \mathbf{x}_{1,t}^* + \boldsymbol{\theta}_2' \mathbf{x}_{2,t}^* + \xi_t \text{ where } \xi_t \underset{c}{\sim} \mathsf{IN}\left[0, \sigma_\xi^2\right], \tag{18}$$

to end with the final selection:

$$y_t = \boldsymbol{\rho}_1' \mathbf{x}_{1,t}^{**} + \boldsymbol{\rho}_2' \mathbf{x}_{2,t}^{**} + \eta_t \text{ where } \eta_t \underset{c}{\sim} \mathsf{IN}\left[0, \sigma_\eta^2\right]. \tag{19}$$

### 8.1 Properties of the selected model

If $\mathbf{x}_{1,t}$ and $\mathbf{x}_{2,t}$ are mutually orthogonal, and (18) is feasible then this procedure delivers the correct answer unless the significance of the relevant variables is close to marginal, so the improved fit of the combination is essential to retain them. Critical values will probably need to be loose in the early subset selections to avoid that problem. Conversely, stringent critical values will probably be needed at the final stage. If, say, $n_1 = n_2 = 100 < T = 150$, then a 1% level would only entail 2 irrelevant variables retained on average despite 200 variables at the start. We first consider the $\mathsf{IID}$ case, so the sub-models are congruent but incomplete.

At stage 1, selecting from (14) when the DGP is in (12), under orthogonality:

$$u_{1,t} = \boldsymbol{\beta}_2' \mathbf{x}_{2,t} + \epsilon_t \tag{20}$$

so $\boldsymbol{\gamma}_1$ is unbiasedly estimated, but with the equation error variance of $\sigma_\epsilon^2 + \boldsymbol{\beta}_2' \mathbf{M}_{2,2} \boldsymbol{\beta}_2$ under stationarity, where $\mathbf{M}_{2,2} = \mathsf{E}\left[\mathbf{x}_{2,t}\mathbf{x}_{2,t}'\right]$. Thus, the primary problems are lack of significance of variables that matter due to 'underfitting', and retention of:

$$\alpha\left(n_1 - k_1\right)$$

irrelevant variables on average when a test of size $\alpha$ is used. Here, we imagine $\alpha = 0.1$ at stage 1, to minimize the loss of variables that matter. For example, if $n_1 = 100$ and $k_1 = 10$ (say), all relevant variables with t-values in excess of about 1.65 in absolute value will be retained together with 9 irrelevant. Similarly for selecting from the $\mathbf{x}_{2,t}$, leading to about 40 variables in the combation of the terminal models:

$$y_t = \boldsymbol{\theta}_1' \mathbf{x}_{1,t}^* + \boldsymbol{\theta}_2' \mathbf{x}_{2,t}^* + \xi_t.$$

At stage 2, set $\alpha = 0.01$ (say), so only about 2 adventitiously-significant variables will on average be retained from the initial 200, whereas all relevant variables that have absolute t-values in excess of about 2.6 in the DGP will be retained. Alternatively, depending on the investigators loss function, $\alpha = 0.025$ would be closer to the value implicit in BIC, and retain variables with absolute t-values in excess of about 2.25.

The third stage may be unnecessary for orthogonal variables, but even there, cross-matching may deliver additional relevant variables in some terminal models, so could be beneficial.

If $\mathbf{x}_{1,t}$ and $\mathbf{x}_{2,t}$ are positively correlated, the efficiency of selection is lower even if the analysis can be conducted in a single stage, and hence must be lower for the multi-stage process proposed here. Nevertheless, we can see that it is likely to work quite well, since the intercorrelations should entail that proxy variables improve fit at the intermediate stages, so could raise the probability of retaining the relevant variables within each subset. However, when the 'correct' regressors are also included, the proxies should be eliminated.

The difficult case is if $\mathbf{x}_{1,t}$ and $\mathbf{x}_{2,t}$ are negatively correlated, since then each is needed for the other to be included. In practice, some negative correlations are likely in amongst some near orthogonal and some positive, so the cross-matching is needed to ensure appropriate pairs at least are always jointly included.

When the data are not IID, so sub-models may be non-congruent, HAC coefficient standard errors may be useful during intermediate stages to ensure that terminal models include all DGP-relevant variables, but should not be needed at the final selection from (18).

## 9 Selecting non-linear models

A relatively common approach in a non-linear setting (see Granger and Teräsvirta, 1993) is to fit non-linear models beginning from a previously selected linear. Such an approach is analogous to simple to general in two respects. First, moves between studies are almost bound to be simple to general, which has poor properties—and may be why empirical advances are so difficult. Secondly, however, any extension of a model should commence from a more general exemplar than the best selected earlier representative, otherwise inbuilt restrictions can preclude finding the appropriate generalization.

Instead, commence with a very general approximation to the non-linearity (such as a polynomial or hypergeometric function, which needs to be identified). Add in the proposed logistic, squashing or whatever functions one at a time and test if they explain the non-linear components of the approximation. This approach avoids the lack of identification under the null, and also directly tests that the postulated functions are the correct ones.

## 10 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 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.

However, 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 using (say) the Liberal strategy, than a researcher commencing from the DGP but selecting (say) the Conservative strategy. Such a finding would have seemed astonishing in the aftermath of Lovell (1983), and both shows the progress and serves to emphasize the importance of the choice of strategy for the underlying selection problem.

The sub-sample reliability procedure appears in Monte Carlo studies to reduce size at a small cost in power, but as yet we have not proved that the resulting trade-off is genuinely beneficial, although it certainly seems relatively costless. Similarly, 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, and briefly described its usage.

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). Thus, we remain confident that further developments will continue to improve the performance of, and widen the scope of application for, automatic modelling procedures.

# 11 Appendix: Progress details

Three sets of experiments are recorded here, re-running Hoover and Perez (1999), Krolzig and Hendry (2001), and Hendry and Krolzig (2002).

### 11.1 The 'Data Mining' experiments re-visited

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:

$$y_t = \delta + \sum_{j=1}^{4} \alpha_j y_{t-j} + \sum_{i=1}^{18} \sum_{j=0}^{1} \gamma_{i,j} x_{i,t-j} + \omega_t. \tag{21}$$

He found none did even reasonably, but in retrospect, that seems mainly because of flaws in the search algorithms evaluated, not the principle of selection *per se*.

Moreover, 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 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 $\alpha = 0.001$ 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 for a single such variable at $\alpha = 0.05$, and about 5% at $\alpha = 0.01$. 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 **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 to date, the DGP is a special case of the GUM, so mis-specification tests play no useful role.

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), and discussed above. *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 be selected.

Table 10 records the DGPs in those experiments which 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.

First, we record the original outcomes reported for our re-run of the Hoover–Perez experiments, shown in table 12. 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 when $\alpha = 0.05$.

**Table 10** 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; HP2* added by the authors.

**Table 11** 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 |

Next, we show how the latest version of *PcGets* (May, 2002) would perform using both Conservative and Liberal settings: see table 13. The outcomes are based on $M = 1000$ replications of the DGP with a sample size of $T = 100$.

**Table 12** 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 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 often above 80% for the Conservative strategy. 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 1% (5% when using the Liberal strategy) – with more than 37 irrelevant variables, the Conservative strategy is clearly the better choice.

The reliability measure is denoted (rel): the size is clearly reduced, being everywhere less than 1% (5%), with little loss of power, confirming the practical value of the reliability check.

The non-deletion and non-selection probabilities are also 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, about 10% for Conservative but above 50% for Liberal, whereas the latter is usually under 5%. Thus, the operating characteristics are stable between the experiments, and quite well behaved.

Overall, these finding cohere with those reported earlier (for a different version of the program, and different settings for the significance levels), and suggest that *PcGets* performs well even in a demanding problem, where the GUM is highly over-parameterized. The outcomes also suggest that relatively loose critical values should be chosen for pre-selection tests.

**Table 13** Re-running the Hoover–Perez experiments.

| Experiment | HP1 | HP2 | HP2* | HP7 |
|---|---|---|---|---|
| *conservative* | | | | |
| T:DGPfound | 1.0000 | 1.0000 | 0.9940 | 1.0000 |
| S:DGPfound | 0.8780 | 0.8290 | 0.6120 | 0.8450 |
| S:NonDeletion | 0.1220 | 0.1700 | 0.2440 | 0.1550 |
| S:NonSelection | 0.0000 | 0.0040 | 0.2360 | 0.0020 |
| T:Dominated | 0.1000 | 0.1160 | 0.1120 | 0.1040 |
| S:Dominated | 0.0220 | 0.0520 | 0.1870 | 0.0490 |
| S:Size | 0.0057 | 0.0091 | 0.0191 | 0.0089 |
| S:Power | —— | 0.9960 | 0.7640 | 0.9987 |
| *reliability based* | | | | |
| S:Size | 0.0037 | 0.0052 | 0.0089 | 0.0054 |
| S:Power | —— | 0.9960 | 0.7629 | 0.9987 |
| *liberal* | | | | |
| T:DGPfound | 1.0000 | 1.0000 | 0.9990 | 1.0000 |
| S:DGPfound | 0.4580 | 0.3390 | 0.3110 | 0.3640 |
| S:NonDeletion | 0.5420 | 0.6610 | 0.6810 | 0.6360 |
| S:NonSelection | 0.0000 | 0.0000 | 0.0260 | 0.0020 |
| T:Dominated | 0.5170 | 0.6170 | 0.6200 | 0.5880 |
| S:Dominated | 0.0250 | 0.0440 | 0.0530 | 0.0460 |
| S:Size | 0.0410 | 0.0538 | 0.0546 | 0.0507 |
| S:Power | —— | 1.0000 | 0.9740 | 0.9993 |
| *reliability based* | | | | |
| S:Size | 0.0354 | 0.0442 | 0.0454 | 0.0422 |
| S:Power | —— | 1.0000 | 0.9725 | 0.9991 |

### 11.2 Re-running the JEDC experiments

In this 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:

$$\begin{aligned} y_t &= \sum_{k=1}^{5} \beta_{k,0} x_{k,t} + \varepsilon_t, & \varepsilon_t &\sim \mathsf{IN}\left[0,1\right], \\ x_t &= v_t, & v_t &\sim \mathsf{IN}_{10}\left[\mathbf{0}, \mathbf{I}_{10}\right] \quad \text{for } t = 1, \ldots, T, \end{aligned} \tag{22}$$

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:

$$y_t = \pi_{0,1} y_{t-1} + \sum_{k=1}^{10} \sum_{i=0}^{1} \pi_{k,i} x_{k,t-i} + \pi_{0,0} + u_t, \;\; u_t \sim \mathsf{IN}\left[0, \sigma^2\right]. \tag{23}$$

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 (23), 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 'canned' 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* (so 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.

### 11.3 Re-running the Stigum experiments

This is the final set we consider for completeness. As the basis for calibrating the current strategies, we simply record *PcGets*' actual performance, since there are no major developments against which to judge progress. The key features are the excellent performance of the Conservative strategy—doing almost as well from the GUM as from the DGP; the accuracy of the actual size for the desired nominal after the reliability check; the small loss of power induced by many irrelevant variables; the rapid reduction in selection error as t-values increase; and the low probabilities of locating the DGP when there are many relevant variables but with t-values around 2 or 3.

**Table 14** JEDC Experiments: Conservative Strategy.

| | Theory | JEDC | JEDC | PcGets | PcGets | PcGets |
|---|---|---|---|---|---|---|
| presearch | | —— | yes | —— | yes | yes |
| spit sample | | —— | —— | —— | —— | yes |
| **Size** | 0.0100 | 0.0185 | 0.0088 | 0.0134 | 0.0106 | 0.0072 |
| vs. JEDC | | | -0.0097 | -0.0051 | -0.0079 | -0.0113 |
| *vs. JEDC (%)* | | | *-52.43* | *-27.57%* | *-42.70%* | *-61.08%* |
| **Power** | 0.7544 | 0.7598 | 0.6665 | 0.7328 | 0.7592 | 0.7412 |
| vs. JEDC | | | -0.0933 | -0.0270 | -0.0006 | -0.0186 |
| *vs. JEDC (%)* | | | *-12.28* | *-3.55%* | *-0.08%* | *-2.45%* |
| **Selection error** | 0.0635 | 0.0689 | 0.0826 | 0.0711 | 0.0629 | 0.0644 |
| vs. JEDC | | | 0.0137 | 0.0022 | -0.0060 | -0.0045 |
| *vs. JEDC (%)* | | | *19.89* | *3.19%* | *-8.66%* | *-6.54%* |
| Costs of search | | 0.0053 | 0.0190 | 0.0075 | -0.0006 | 0.0008 |
| *Costs of search (%)* | | *8.40* | *29.97* | *11.86%* | *-0.99%* | *1.32%* |
| **Expected number of** | | | | | | |
| NonDGP variables incl. | 0.17 | 0.31 | 0.15 | 0.23 | 0.18 | 0.12 |
| DGP variables deleted | 1.23 | 1.20 | 1.67 | 1.34 | 1.20 | 1.29 |
| variables misplaced | 1.40 | 1.52 | 1.82 | 1.56 | 1.38 | 1.42 |
| **Power function** | | | | | | |
| power (t=2) | 0.2580 | 0.2820 | 0.1538 | 0.2370 | 0.2880 | 0.2560 |
| power (t=3) | 0.6130 | 0.6200 | 0.4278 | 0.5730 | 0.6230 | 0.5854 |
| power (t=4) | 0.9020 | 0.8980 | 0.7645 | 0.8540 | 0.8860 | 0.8668 |
| power (t=6) | 0.9990 | 0.9990 | 0.9865 | 1.0000 | 0.9990 | 0.9981 |
| 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 | | | | 0.1000 | 0.1430 | 0.1430 |
| S:NonDeletion | | | | 0.1790 | 0.1330 | 0.1330 |
| S:NonSelection | | | | 0.8750 | 0.8410 | 0.8410 |
| T:Dominated | | | | 0.6890 | 0.7460 | 0.7460 |
| S:Dominated | | | | 0.1500 | 0.0680 | 0.0680 |

**Table 15** JEDC Experiments: Liberal Strategy.

| | Theory | JEDC | JEDC | PcGets | PcGets | PcGets |
|---|---|---|---|---|---|---|
| presearch | | —— | yes | —— | yes | yes |
| spit sample | | —— | —— | —— | —— | yes |
| **Size** | 0.0500 | 0.0677 | 0.0477 | 0.0658 | 0.0546 | 0.0486 |
| vs. JEDC | | | -0.0200 | -0.0019 | -0.0131 | -0.0191 |
| *vs. JEDC (%)* | | | *-29.54* | *-2.81%* | *-19.36%* | *-28.21%* |
| **Power** | 0.8522 | 0.8532 | 0.8156 | 0.8556 | 0.8446 | 0.8350 |
| vs. JEDC | | | -0.0376 | 0.0024 | -0.0086 | -0.0182 |
| *vs. JEDC (%)* | | | *-4.41* | *0.28%* | *-1.01%* | *-2.13%* |
| **Selection error** | 0.0722 | 0.0857 | 0.0788 | 0.0837 | 0.0775 | 0.0751 |
| vs. JEDC | | | -0.0069 | -0.0020 | -0.0082 | -0.0106 |
| *vs. JEDC (%)* | | | *-8.06* | *-2.35%* | *-9.54%* | *-12.40%* |
| Costs of search | | 0.0135 | 0.0065 | 0.0114 | 0.0053 | 0.0028 |
| *Costs of search (%)* | | *18.62* | *9.06* | *15.83%* | *7.30%* | *3.91%* |
| **Expected number of** | | | | | | |
| NonDGP variables incl. | 0.85 | 1.15 | 0.81 | 1.12 | 0.93 | 0.83 |
| DGP variables deleted | 0.74 | 0.73 | 0.92 | 0.72 | 0.78 | 0.83 |
| variables misplaced | 1.59 | 1.88 | 1.73 | 1.84 | 1.71 | 1.65 |
| **Power function** | | | | | | |
| power (t=2) | 0.4730 | 0.4930 | 0.4080 | 0.4840 | 0.4890 | 0.4482 |
| power (t=3) | 0.8120 | 0.8020 | 0.7330 | 0.8120 | 0.8140 | 0.7813 |
| power (t=4) | 0.9760 | 0.9720 | 0.9390 | 0.9600 | 0.9600 | 0.9453 |
| 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 | 0.1810 | 0.1810 |
| S:NonDeletion | | | | 0.6290 | 0.5620 | 0.5620 |
| S:NonSelection | | | | 0.5960 | 0.6140 | 0.6140 |
| T:Dominated | | | | 0.7550 | 0.7270 | 0.7270 |
| S:Dominated | | | | 0.0470 | 0.0230 | 0.0230 |

**Table 16** Liberal and Conservative strategies.

| *DGP Design* | A | B | C | D | E | F | G | H |
|---|---|---|---|---|---|---|---|---|
| t | 0 | 0 | 2 | 2 | 3 | 3 | 4 | 4 |
| $k$ | 8 | 8 | 8 | 8 | 8 | 8 | 8 | 8 |
| $n$ | 8 | 20 | 8 | 20 | 8 | 20 | 8 | 20 |
| *conservative* | probabilities | | | | | | | |
| T:DGPfound | 1.0000 | 1.0000 | 0.0000 | 0.0000 | 0.0230 | 0.0280 | 0.3440 | 0.3630 |
| S:DGPfound | 0.8590 | 0.8080 | 0.0000 | 0.0000 | 0.0290 | 0.0250 | 0.3420 | 0.2780 |
| S:NonDeletion | 0.1410 | 0.1920 | 0.1010 | 0.2260 | 0.1000 | 0.2310 | 0.0840 | 0.2000 |
| S:NonSelection | 0.0000 | 0.0000 | 0.9990 | 0.9980 | 0.9660 | 0.9660 | 0.6350 | 0.6530 |
| T:Dominated | 0.1300 | 0.1750 | 0.8560 | 0.6370 | 0.7740 | 0.4920 | 0.5620 | 0.4320 |
| S:Dominated | 0.0110 | 0.0170 | 0.0840 | 0.2270 | 0.1380 | 0.3300 | 0.0640 | 0.2090 |
| S:Size | 0.0104 | 0.0095 | 0.0129 | 0.0137 | 0.0127 | 0.0153 | 0.0098 | 0.0137 |
| S:Power | —— | —— | 0.3250 | 0.2893 | 0.6609 | 0.6108 | 0.8854 | 0.8578 |
| S:Selection error | —— | —— | 0.2973 | 0.2129 | 0.1759 | 0.1222 | 0.0622 | 0.0504 |
| *reliability based* | | | | | | | | |
| S:Size | 0.0075 | 0.0069 | 0.0084 | 0.0093 | 0.0086 | 0.0106 | 0.0069 | 0.0083 |
| S:Power | —— | —— | 0.2562 | 0.2446 | 0.5786 | 0.5571 | 0.8454 | 0.8309 |
| S:Selection error | —— | —— | 0.2973 | 0.2225 | 0.2150 | 0.1341 | 0.0808 | 0.0542 |
| *liberal* | probabilities | | | | | | | |
| T:DGPfound | 1.0000 | 1.0000 | 0.0010 | 0.0050 | 0.1740 | 0.2000 | 0.7330 | 0.7310 |
| S:DGPfound | 0.4030 | 0.3290 | 0.0020 | 0.0030 | 0.1200 | 0.0770 | 0.4520 | 0.2670 |
| S:NonDeletion | 0.5970 | 0.6710 | 0.4340 | 0.6560 | 0.4370 | 0.6390 | 0.4010 | 0.6170 |
| S:NonSelection | 0.0000 | 0.0000 | 0.9950 | 0.9950 | 0.8160 | 0.8270 | 0.2640 | 0.3400 |
| T:Dominated | 0.5660 | 0.6470 | 0.8640 | 0.7930 | 0.7310 | 0.6900 | 0.4940 | 0.5980 |
| S:Dominated | 0.0310 | 0.0240 | 0.0320 | 0.0290 | 0.0500 | 0.0440 | 0.0270 | 0.0390 |
| S:Size | 0.0548 | 0.0482 | 0.0655 | 0.0593 | 0.0644 | 0.0595 | 0.0569 | 0.0553 |
| S:Power | —— | —— | 0.4933 | 0.4765 | 0.8001 | 0.7789 | 0.9600 | 0.9466 |
| S:Selection error | —— | —— | 0.2973 | 0.1919 | 0.1321 | 0.1056 | 0.0485 | 0.0547 |
| *reliability based* | | | | | | | | |
| S:Size | 0.0450 | 0.0407 | 0.0539 | 0.0497 | 0.0523 | 0.0498 | 0.0462 | 0.0459 |
| S:Power | —— | —— | 0.4457 | 0.4389 | 0.7608 | 0.7466 | 0.9426 | 0.9324 |
| S:Selection error | —— | —— | 0.2973 | 0.1958 | 0.1458 | 0.1080 | 0.0518 | 0.0521 |

## References


Akaike, A. (1973). Information theory and an extension of the maximum likelihood principle. In Petrov, B. N., and Saki, F. L. (eds.), *Second International Symposium of Information Theory*. Budapest.

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

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

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.

Brüggemann, R., Krolzig, H.-M., and Lütkepohl, H. (2002). Comparison of model selection procedures for VAR processes. Mimeo, Humboldt–University, Berlin.

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.

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., *Econometrics: Alchemy or Science?* Oxford: Blackwell Publishers, 1993, and Oxford University Press, 2000.

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

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

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

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

Granger, C. W. J., and Teräsvirta, T. (1993). *Modelling Nonlinear Economic Relationships*. Oxford: Oxford University Press.

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. (2002). 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.

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.

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

Lynch, A. W., and Vital-Ahuja, T. (1998). Can subsample evidence alleviate the data-snooping problem? a comparison to the maximal $R^2$ cutoff test. Discussion paper, New York University.

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

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.

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.