

# Consistent Model Selection by an Automatic *Gets* Approach

JULIA CAMPOS, DAVID F. HENDRY and HANS-MARTIN KROLZIG\*

Economics Department, University of Salamanca, and  
Economics Department, Oxford University, Manor Road, Oxford OX1 3UQ.  
(e-mail: david.hendry@nuffield.ox.ac.uk)

October 19, 2003

## Abstract

We establish the consistency of the selection procedures embodied in *PcGets*, and compare their performance with other model selection criteria in linear regressions. The significance levels embedded in the *PcGets* Liberal and Conservative algorithms coincide in very large samples with those implicit in the Hannan–Quinn (*HQ*) and Schwarz information criteria (*SIC*) respectively. Thus, both *PcGets* rules are consistent under the same conditions as *HQ* and *SIC*. However, *PcGets* has rather different finite-sample behaviour. Pre-selecting to remove many of the candidate variables is confirmed as enhancing the performance of *SIC*.

## 1 Introduction

Econometric model selection is a venerable problem, for which many different solutions have been proposed. Recent advances in computer automation of general-to-specific (*Gets*) methods have thrown fresh light on the potential of that approach, both by revealing some high success rates, and by allowing operational studies of alternative tactics: see *inter alia* Hoover and Perez (1999) and Krolzig and Hendry (2001). An overview of the literature, and the developments leading to *Gets* modelling in particular, is provided by Campos, Ericsson and Hendry (2003).

Hendry and Krolzig (2001) describe the selection algorithms embodied in *PcGets*, their foundation in the theory of reduction, and potential alternatives.<sup>1</sup> There are two pre-programmed procedures, called the Liberal and Conservative strategies: the former seeks a null rejection frequency per candidate variable in a regression of about 5%, whereas the latter is centered on 1%. The Monte Carlo evidence in Hendry and Krolzig (2003a) on the performance of *PcGets* across a range of experiments confirms that these strategies have their intended null rejection frequencies, and given those, both have power close to the optimum obtainable, namely that of an equivalent significance level single test when the form of the distribution is known. Finally, Hendry and Krolzig (2003b) show how to produce nearly unbiased estimates despite selection, with estimated standard errors close to those found for the estimated (correctly specified) equation in the data generation process (DGP). In this paper, we establish the consistency of the two main selection strategies embodied in *PcGets*: for a related analysis, see White (1990).

---

\*We are grateful to Oxford University Research Development Fund for financial support, and to Bent Nielsen and Peter Phillips for helpful comments on an earlier draft.

<sup>1</sup>*PcGets* is an Ox Package (see Doornik, 1999) implementing automatic *Gets* modelling for linear regression models, based on the theory of reduction: see e.g., Hendry (1995, Ch.9).

The paper is organized as follows. First, section 2 formulates the setting. Then, section 3 maps the selection rules of the Hannan–Quinn criterion (denoted  $HQ$ : see Hannan and Quinn, 1979), and the Schwarz information criterion ( $SIC$ : see Schwarz, 1978) into implicit significance levels. Next, using that mapping, section 4 considers the consistency of the two *PcGets* model selection strategies. Section 5 compares the finite-sample performance of *PcGets* with  $SIC$ . Section 6 concludes.

## 2 Formulation

Consider a general unrestricted model (GUM) of  $y_t$  where there are  $n$  candidate regressor variables  $\mathbf{z}'_t = (z_1 \dots z_n)$  over a sample  $t = 1, \dots, T$ :

$$y_t = \sum_{i=1}^n \gamma_i z_{i,t} + v_t \text{ where } v_t \sim \text{IN} [0, \sigma_v^2] \quad (1)$$

when the GUM in (1) is fully congruent, so matches the data evidence in all relevant respects. Letting  $E[\cdot]$  denote an expectation, then  $E[\mathbf{z}_t v_t] = 0$ ; and if  $\mathbf{y}' = (y_1 \dots y_T)$  and  $\mathbf{Z}' = (\mathbf{z}_1 \dots \mathbf{z}_T)$ :

$$\hat{\gamma} = (\mathbf{Z}'\mathbf{Z})^{-1} \mathbf{Z}'\mathbf{y} \quad (2)$$

which satisfies  $E[\hat{\gamma}] = \gamma$ . Also, when  $V[\cdot]$  denotes a variance:

$$V[\hat{\gamma}] = \sigma_v^2 (\mathbf{Z}'\mathbf{Z})^{-1} \quad (3)$$

and for  $\hat{v}_t = y_t - \hat{\gamma}'\mathbf{z}_t$ :

$$\hat{\sigma}_v^2 = \frac{\sum_{t=1}^T \hat{v}_t^2}{T - n} \quad (4)$$

where  $E[\hat{\sigma}_v^2] = \sigma_v^2$ . Thus, the estimates in the GUM are unbiased, but inefficient when some of the  $\gamma_i$  are zero.

The DGP equation only involves  $m \leq n$  variables:

$$y_t = \sum_{j=1}^m \beta_j z_{j,t} + \epsilon_t \text{ where } \epsilon_t \sim \text{IN} [0, \sigma_\epsilon^2] \quad (5)$$

so the DGP is nested in the GUM (1). For convenience, (5) is written as if the first  $m \leq n$  regressors are assumed to be the relevant ones, but the investigator does not know that information. Similar formulae to (2), (3) and (4) hold for the DGP specification, since  $\gamma$  is  $\beta$  augmented by zeros:

$$\gamma = \begin{pmatrix} \beta \\ \mathbf{0} \end{pmatrix}.$$

### 2.1 Asymptotic behaviour

We assume the properties of  $\mathbf{Z}$  are such that:

$$\text{plim}_{T \rightarrow \infty} T^{-1} \mathbf{Z}'\mathbf{Z} = \mathbf{Q}_{zz}, \quad (6)$$

where  $\mathbf{Q}_{zz}$  is finite positive definite. Generalizations to integrated processes seem feasible, but would require different normalization factors. Then given (1), conditional on  $\mathbf{Z}$ :

$$\sqrt{T}(\hat{\gamma} - \gamma) \xrightarrow{D} \mathbf{N}[\mathbf{0}, \sigma_v^2 \mathbf{Q}_{zz}^{-1}]. \quad (7)$$

Estimation of (1) also delivers a consistent estimator  $\hat{\sigma}_v^2$  of  $\sigma_v^2$ .

Letting  $h^{ii}$  denote the  $i^{th}$  diagonal element of  $(\mathbf{Z}'\mathbf{Z})^{-1}$ , the individual rescaled  $t^2$ -statistics for each  $\gamma_i$  are:

$$T^{-1}t_{\gamma_i}^2 = \frac{T^{-1}\hat{\gamma}_i^2}{\hat{\sigma}_v^2 h^{ii}} \simeq \frac{\hat{\gamma}_i^2}{\hat{\sigma}_v^2 q^{ii}}, \quad (8)$$

where  $q^{ii} > 0$  is the  $i^{th}$  diagonal element of  $\mathbf{Q}_{zz}^{-1}$ . Then if  $\gamma_i = \beta_i \neq 0$ :

$$\text{plim}_{T \rightarrow \infty} T^{-1}t_{\gamma_i}^2 = \text{plim}_{T \rightarrow \infty} \frac{\hat{\gamma}_i^2}{\hat{\sigma}_v^2 q^{ii}} = \frac{\beta_i^2}{\sigma_v^2 q^{ii}}, \quad (9)$$

so:

$$\text{plim}_{T \rightarrow \infty} t_{\gamma_i}^2 = \text{plim}_{T \rightarrow \infty} \frac{T\beta_i^2}{\sigma_v^2 q^{ii}} \rightarrow \infty \text{ when } \beta_i^2 \neq 0. \quad (10)$$

Thus, however small the significance level  $\alpha \neq 0$ , or large, but finite, the critical value  $c_\alpha > 0$ , where:

$$\mathbf{P}\left(t_{\gamma_i}^2 > c_\alpha \mid \beta_i^2 = 0\right) = \alpha, \quad (11)$$

then  $t_{\gamma_i}^2$  will reject a false null hypothesis  $\gamma_i = 0$  with probability unity:

$$\mathbf{P}\left(t_{\gamma_i}^2 > c_\alpha \mid \beta_i^2 \neq 0\right) \rightarrow 1. \quad (12)$$

Conversely, if  $\gamma_i = 0$  in (1), then:

$$\text{plim}_{T \rightarrow \infty} T^{-1}t_{\gamma_i}^2 \rightarrow 0, \quad (13)$$

and indeed  $t_{\gamma_i}^2$  has an asymptotic central  $F_{T-n}^1$  distribution. Thus, for a sufficiently large finite critical value  $c_\alpha > 0$  which increases at a suitable rate with  $T$ :

$$\mathbf{P}\left(t_{\gamma_i}^2 \leq c_\alpha \mid \gamma_i^2 = 0\right) \rightarrow 1. \quad (14)$$

Hence, for finite  $n > m$  and  $T \rightarrow \infty$ , the GUM provides consistent estimates, and with an appropriate sequence of  $c_\alpha > 0$ , consistent tests. We now consider the rate of increase of  $c_\alpha > 0$  with  $T$ , drawing on the literature showing the consistency of information criteria selection.

## 2.2 Information criteria

The Schwarz (1978) information criterion, *SIC* (independently derived by Rissanen, 1978, using coding theory) selects from the set of  $n$  candidate variables in (1) the model with  $k$  regressors and parameter estimates  $\tilde{\delta}_i$  which minimizes:

$$SIC_k = \ln \tilde{\sigma}_k^2 + c \frac{k \ln T}{T}. \quad (15)$$

Here,  $c \geq 1$  was introduced by Hannan and Quinn (1979) (and was unity originally) with:

$$\tilde{\sigma}_k^2 = \frac{1}{T} \sum_{t=1}^T \left( y_t - \sum_{i=1}^k \tilde{\delta}_i z_{i,t} \right)^2 = \frac{1}{T} \sum_{t=1}^T \tilde{u}_t^2 \quad (16)$$

where (again assuming an appropriate ordering of the variables):

$$\tilde{u}_t = y_t - \sum_{i=1}^k \tilde{\delta}_i z_{i,t}. \quad (17)$$

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  $10^{12}$ . The  $HQ$  criterion replaces the last term of (15) by  $2k \ln(\ln T)/T$ .

One of the first information criteria for model selection was proposed by Akaike (1969), namely  $FPE$  (for final prediction error), followed in Akaike (1973) by  $AIC$ , which penalizes the log-likelihood by  $2k/T$  for  $k$  parameters and a sample size of  $T$ :<sup>2</sup>

$$AIC_k = \ln \tilde{\sigma}_k^2 + \frac{2k}{T}. \quad (18)$$

Shibata (1980) showed that  $AIC$  is an asymptotically efficient selection method when the DGP is an infinite order process. However, as Hannan and Quinn (1979) show,  $AIC$  does not guarantee a consistent selection as the sample size diverges when the DGP is nested in the GUM. Nevertheless, Sober (2003) argues that there are good philosophical grounds for its adoption in an instrumentalist approach, where prediction is pre-eminent.

Finally, Phillips (1994, 1995, 1996) has proposed an automated model selection approach based on a posterior information criterion for forecasting ( $PICF$ ), which re-selects the specification of the model and re-estimates the resulting parameters (including e.g., cointegration rank) as new information accrues. He establishes its properties under general conditions for non-stationary, evolving processes where the DGP need not be nested in the GUM (also see Phillips and Ploberger, 1995, who propose  $PIC$  as an extension of  $SIC$ ).

### 2.3 *PcGets*

Let  $\mathcal{S}_r$  and  $\mathcal{S}_0$  respectively denote the sets of retained relevant and irrelevant variables, so the model selected by *PcGets* is:

$$y_t = \sum_{i \in \mathcal{S}_r} \theta_i z_{i,t} + \sum_{j \in \mathcal{S}_0} \rho_j z_{j,t} + w_t \quad (19)$$

where  $\mathcal{S}_r$  has  $p \leq m$  elements and  $\mathcal{S}_0$  has  $q \leq (n - m)$ .

## 3 Mapping information criteria to significance levels

First, we formally establish the well-known link of  $SIC$  to significance levels. We then note the implicit setting of significance levels involved in the choice of  $c$  in (15), record the corresponding formulae for  $HQ$ , and note the potential role of  $AIC$ .

Consider the impact on (15) of adding an extra orthogonalized regressor  $z_{k+1,t}$  to the linear regression model with  $k$  such variables, and residuals given by (17). Orthogonality is convenient for simplifying the proof, but not essential. Then:

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

---

<sup>2</sup>Perhaps the oldest model selection criterion is the unbiased residual variance criterion proposed by Theil (1961).

Thus, as is well known, from (16):

$$\begin{aligned}
\tilde{\sigma}_{k+1}^2 &= \frac{1}{T} \sum_{t=1}^T \left( \tilde{u}_t - \hat{\beta}_{k+1} z_{k+1,t} \right)^2 \\
&= \tilde{\sigma}_k^2 \left( 1 - \frac{\hat{\beta}_{k+1}^2 \sum_{t=1}^T z_{k+1,t}^2}{T \tilde{\sigma}_k^2} \right) \\
&= \tilde{\sigma}_k^2 \left( 1 - (T - k - 1)^{-1} \hat{\mathbf{t}}_{(k+1)}^2 \frac{\tilde{\sigma}_{k+1}^2}{\tilde{\sigma}_k^2} \right)
\end{aligned} \tag{20}$$

so that:

$$\tilde{\sigma}_{k+1}^2 = \tilde{\sigma}_k^2 \left( 1 + (T - k - 1)^{-1} \hat{\mathbf{t}}_{(k+1)}^2 \right)^{-1}. \tag{21}$$

In (20):

$$\hat{\mathbf{t}}_{(k+1)}^2 = \frac{T \hat{\beta}_{k+1}^2 \sum_{t=1}^T z_{k+1,t}^2}{\hat{\sigma}_{k+1}^2} \tag{22}$$

is the square of the conventional **t**-test of the null hypothesis that  $\beta_{k+1} = 0$ . The subscript in parentheses on **t** in (22) denotes the marginal regressor under consideration: if required, the degrees of freedom could be shown as, e.g.,  $\hat{\mathbf{t}}_{(k+1)}^2 (T - k - 1)$ . Also,  $\hat{\sigma}_{k+1}^2$  is an unbiased estimator of the error variance:

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

Consequently, from (15):

$$\begin{aligned}
SIC_{k+1} &= \ln \tilde{\sigma}_{k+1}^2 + c \frac{(k+1) \ln T}{T} \\
&= \ln \tilde{\sigma}_k^2 + c \frac{k \ln T}{T} - \ln \left( 1 + (T - k - 1)^{-1} \hat{\mathbf{t}}_{(k+1)}^2 \right) + c \frac{\ln T}{T} \\
&= SIC_k + \frac{c}{T} \ln T - \ln \left( 1 + (T - k - 1)^{-1} \hat{\mathbf{t}}_{(k+1)}^2 \right).
\end{aligned} \tag{23}$$

Hence,  $SIC_{k+1} < SIC_k$  when:

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

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

$$\hat{\mathbf{t}}_{(k+1)}^2 > (T - k - 1) \left( T^{c/T} - 1 \right). \tag{24}$$

Thus, given a value for  $c$ , (24) determines the implicit significance level of  $SIC$  as a function of  $T$  and  $k$ . For example, when  $T = 140$ , with  $c = 1$  (the usual choice), and  $k = 40$  as in Hoover and Perez (1999), then  $SIC_{41} < SIC_{40}$  whenever  $\hat{\mathbf{t}}_{(41)}^2 \geq 3.56$ , or  $|\mathbf{t}_{(41)}| \geq 1.89$ . This is close to the 6% level.

Choosing  $c > 1$  is tantamount to choosing a more stringent **p**-value for the corresponding **t**-test: e.g., setting  $c = 2$  in (24) for the same  $T$  and  $k$  entails  $SIC_{41} < SIC_{40}$  whenever  $\hat{\mathbf{t}}_{(41)}^2 \geq 7.24$ , or  $|\mathbf{t}_{(41)}| \geq 2.69$ , which now maps into a 0.84% test: section 5 considers in more detail the impact of changing  $c$  in finite samples.

More generally, consider two nested models with  $k_1 < k$  and  $k$  regressors respectively, and error variances  $\tilde{\sigma}_{k_1}^2$  and  $\tilde{\sigma}_k^2$ . The smaller model will be chosen if  $SIC_{k_1} < SIC_k$  where:

$$SIC_{k_1} - SIC_k = \ln \frac{\tilde{\sigma}_{k_1}^2}{\tilde{\sigma}_k^2} - \ln T^{c(k-k_1)/T} < 0.$$

The F-statistic for testing the first against the second is:

$$F_{T-k}^{k-k_1} = \left( \frac{T-k}{k-k_1} \right) \left( \frac{(T-k_1)\tilde{\sigma}_k^2}{(T-k)\tilde{\sigma}_{k_1}^2} - 1 \right),$$

so that the first will be chosen by  $SIC$  if:

$$F_{T-k}^{k-k_1} < \left( \frac{T-k}{k-k_1} \right) \left( \frac{(T-k_1) T^{c(k-k_1)/T}}{T-k} - 1 \right). \quad (25)$$

Similar logic reveals that the  $(k+1)^{st}$  additional regressor will be retained by  $HQ$  when:

$$\hat{t}_{(k+1)}^2 > (T-k-1) \left[ (\ln T)^{1/2T} - 1 \right] \quad (26)$$

which for  $T = 140$  and  $k = 40$  leads to  $\hat{t}_{(41)}^2 \geq 2.29$  or about a 13% significance level. Finally,  $AIC_{k+1} < AIC_k$  when:

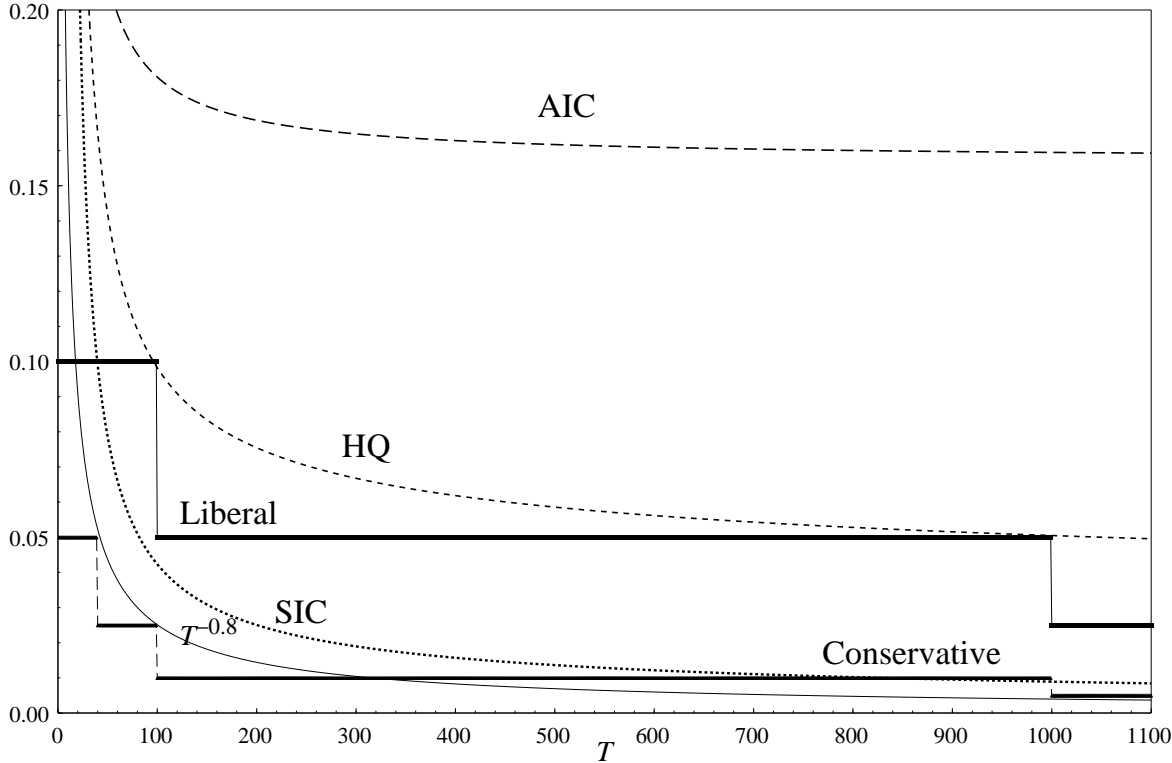
$$\hat{t}_{(k+1)}^2 > (T-k-1) \left( \exp \left( \frac{2}{T} \right) - 1 \right). \quad (27)$$

Now, for  $T = 140$  and  $k = 40$ , (27) just requires  $\hat{t}_{(41)}^2 \geq 1.42$  or about a 24% significance level. The derived profiles of significance levels for all these criteria are shown in figures 1 and 2 below for  $k = 10$  and  $k = 40$ .

## 4 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 e.g., Hannan and Quinn (1979). Although  $AIC$  is not consistent, both  $SIC$  and  $HQ$  are, in that they ensure that a DGP nested within a model thereof will be selected with probability unity as  $T$  diverges relative to  $n$ . Atkinson (1981) proposes a general function from which various criteria for model selection can be generated.

Consistent selection requires that the number of observations per parameter diverges at an appropriate rate, so that non-zero non-centralities increase indefinitely (guaranteeing retention of relevant variables), and that the significance level of the procedure converges towards zero at an appropriate rate, so irrelevant variables are eventually retained with probability zero. In particular,  $SIC$  penalizes the log-likelihood by  $k \ln(T)/T$  as in (15), whereas  $HQ$  uses  $2k \ln(\ln(T))/T$ , which Hannan and Quinn (1979) show is the minimum rate at which additional parameters must be penalized. Then selection is strongly consistent against fixed alternatives when the assumed order  $n$  of the general model is no less than the true order  $m$ , and  $n/T \rightarrow 0$ . Based on a Monte Carlo, Hannan and Quinn (1979) suggest that  $HQ$  may perform better than  $SIC$  in large sample sizes. When  $n$  increases with the sample size but  $n/T \rightarrow 0$ , the large model analysis in Sargan (1975) could be adapted to establish consistent GUM estimates (also see Robinson, 2003, who re-interprets Sargan's approach as a semi-parametric estimator).



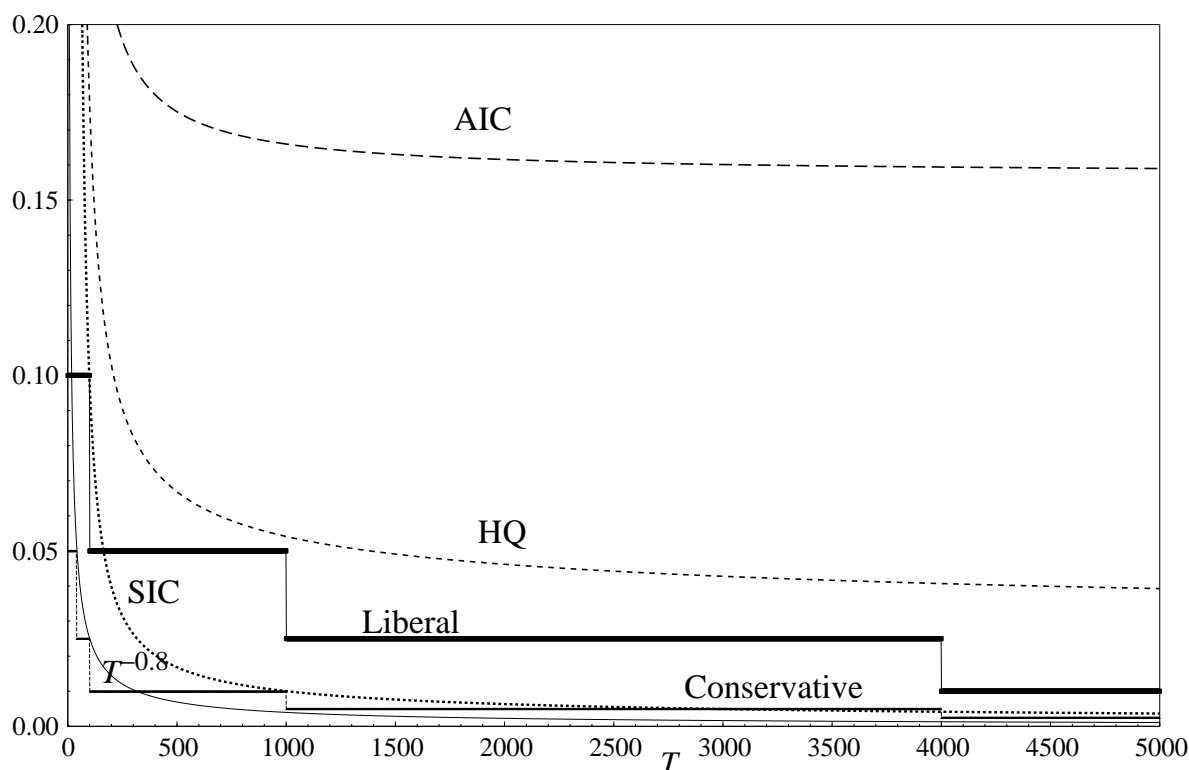
**Figure 1** Significance level comparisons for 10 regressors.

When all variables are mutually orthogonal, the *PcGets* procedure is equivalent to simply ranking the squared t-tests on each variable  $z_i$  in the GUM (1) in section 2, and retaining all and only those  $t_{\gamma_i}^2$  that exceed the pre-set significance level  $c_\alpha$ . Thus, the final model would actually be selected by a single decision.  $t_{\gamma_i}^2$  corresponding to non-zero DGP parameters diverge, whereas all others are distributed as central  $F_{T-n}^1$  so the decision is clear cut. When regressors are not mutually orthogonal, the multi-path search procedure in effect implements a sequence of such transforms, since eliminating variables is equivalent to dropping their orthogonal component relative to the retained variables.

To establish the consistency of the Liberal and Conservative strategies in *PcGets*, we use the mappings in section 3 of *HQ* and *SIC* to implicit significance levels, then show that the two *PcGets* rules converge respectively to these. *PcGets* also has similar general requirements to those needed for consistent selection: the GUM is assumed to be over-parameterized relative to the (local) DGP, and the nominal significance levels tend to zero at an appropriate rate 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 bound, and *SIC* as its lower). The Conservative strategy uses a more stringent significance level, implicitly attributing a higher cost to retaining variables that do not matter in the DGP. Its initial significance level is, therefore, more stringent than *SIC* (for  $c = 1$ ), but converges to *SIC* as the sample size becomes very large.

Figure 1 illustrates the *PcGets* rules for 10 variables relative to *AIC*, *SIC* and *HQ* for sample sizes up to 1100 in the space of entailed significance levels. As can be seen, the *PcGets* Conservative profile is initially much tighter than any of the three information criteria considered, whereas the Liberal strategy usually lies between *HQ* (as its upper bound) and *SIC* (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 imple-

mented with ease, such as that using the nearest selection criterion value, or  $T^{-0.8}$  (also shown, based on Hendry, 1995, Ch. 13). However, as the two pre-programmed strategies are designed for relatively non-expert users, it seems preferable to base them more closely on ‘conventional’ significance levels. *AIC* is substantially less stringent, particularly at larger sample sizes, so would tend to over-select when there are many irrelevant candidate variables. However, the Conservative profile is noticeably tighter than *SIC* at small samples, so the next section compares it with *SIC*. Importantly, while both *SIC* and *HQ* deliver consistent selections, they could differ substantively in small samples, and it is precisely the intent of the two *PcGets* strategies to outperform for models and samples of a size relevant to macro-econometrics. 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, and hence the choice of strategy.



**Figure 2** Significance level comparisons for 40 regressors.

Figure 2 shows the corresponding comparisons for 40 variables to illustrate the impact of increasing that dimension, now for  $T \leq 5000$ . Even at such large sample sizes, the implicit significance levels of the various rules differ substantially. Both these figures hold  $n$  fixed as  $T$  diverges, albeit at very different levels. Figure 3 focuses on smaller sample sizes, namely  $T \leq 200$ , and now compares the two sets ( $n = 10$  and  $n = 40$ ), to show the impact of many more variables. The selection rules have increasingly high levels of significance as the number of candidate variables approaches the sample size (i.e.,  $n \rightarrow T$ ), which does not seem a desirable feature.<sup>3</sup> However, based on Sugiura (1978), Hurvich and Tsai (1989) derive a non-stochastic correction to *AIC* that retains asymptotically efficient selection, but corrects its bias as an estimator of the Kullback–Leibler information discrepancy measure (see e.g., Kullback and Leibler, 1951), when there are large numbers of parameters relative to the available sample. This appears

<sup>3</sup>*PcGets* can handle  $n > T$ , by repeated block searches (see Hendry and Krolzig, 2003c), so maintains relatively tighter levels.



to produce a uniform improvement in its behaviour, and is given by (in our notation):

$$AIC_k^c = AIC_k + \frac{2k(k+1)}{T(T-k-1)}. \quad (28)$$

Even for  $T = 140$  and  $k = 40$ , (28) produces a much tighter significance level of  $\hat{t}_{(k+1)}^2 > 2.61$ , which is about 11%, and more stringent than  $HQ$ . As  $n \rightarrow T$ , models with  $k \simeq n$  will be strongly selected against, removing the problem apparent in figure 3.

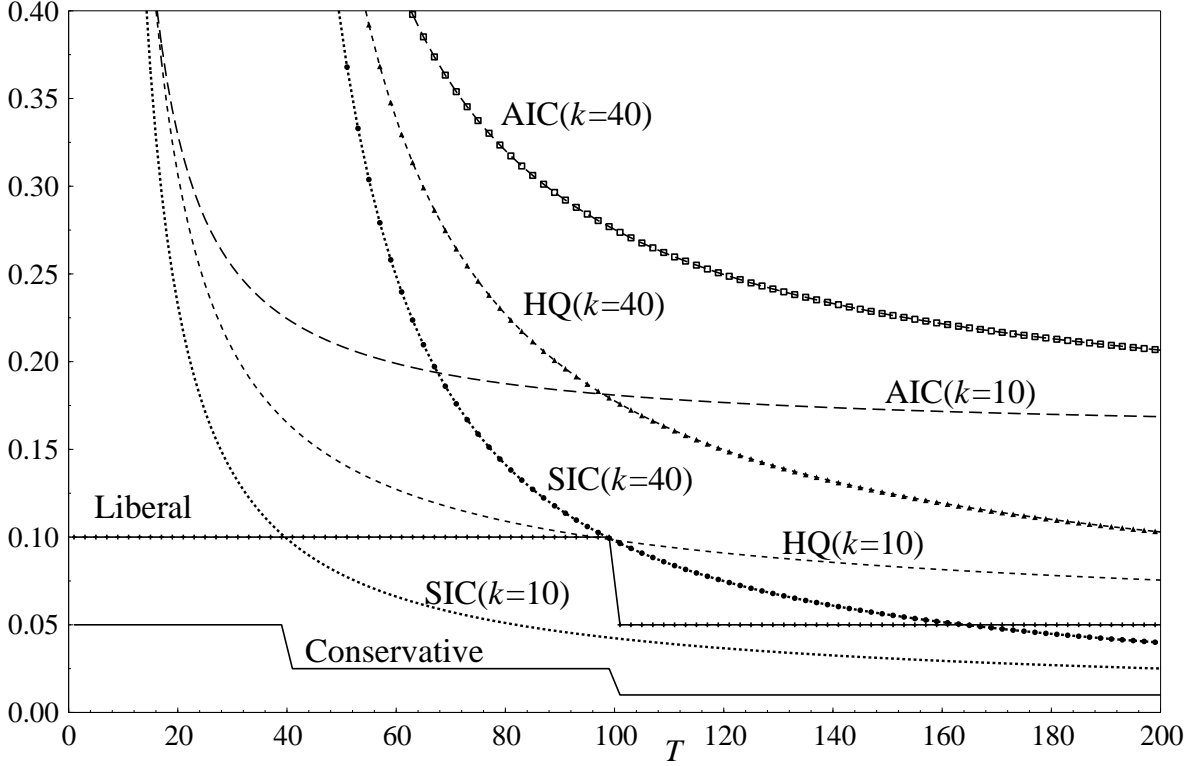


Figure 3 Small sample comparisons for 10 and 40 regressors.

## 5 Comparisons with $SIC$

Having shown in figures 1–3 the effects of altering the form of the penalty function across the various criteria, we now consider the relative behaviour of  $PcGets$  and  $SIC$  when the null model is true. Then we turn to the impact of pre-selection on performance, namely reducing the value of  $n$  to a manageable number of models, to explain the outcomes described by Hansen (1999). To find the DGP by  $SIC$  for  $c = 1$  when the null model ( $H_0$ ) is true for  $T = 140$  requires it to select no variables, so that:

$$\hat{t}_{(k)}^2 \leq (T - k) \left( T^{\frac{1}{T}} - 1 \right), \forall k \leq n \quad (29)$$

which is a sequence of  $\hat{t}_{(i)}^2$  between 3.59 (at  $k = 40$ ) and 4.49 (at  $k = 1$ ). That outcome entails at least every  $\hat{t}_{(i)}^2 < 3.59$ , which has a probability, in an orthogonal setting, even using as an approximation the best case of 140 degrees of freedom (when no candidate variables are left):

$$P \left( \hat{t}_{(i)}^2 \leq 3.59 \forall i = 1, \dots, 40 \mid H_0 \right) = (1 - 0.06)^{40} = 0.08. \quad (30)$$

Thus 92% of the time *SIC* should retain some irrelevant variable(s).

More formally, let the  $k$  mutually independent explanatory variables ordered by their squared  $t$ -values around zero be denoted by  $\tau_i$  so that  $0 \leq \tau_1 \leq \tau_2 \leq \dots \leq \tau_k < \infty$ . We wish to compute the probabilities of exclusion when all  $k$  variables are irrelevant, so the  $t_{(i)}^2$  are distributed as central  $F_{T-k}^1$ . Because the  $F$ -statistics are independent, the joint density of the  $\{\tau_i\}$  is:

$$g_{\tau_1, \tau_2, \dots, \tau_k}(\tau_1, \tau_2, \dots, \tau_k) = k! \prod_{i=1}^k g(\tau_i),$$

where  $g(\tau_i)$  is the density of a central  $F_{T-k}^1$ . For an  $\alpha\%$  significance level, the probability of correctly excluding all  $k$  variables is:

$$\begin{aligned} & P(\tau_i \leq c_\alpha, i = 1, \dots, k \mid H_0) \\ &= P(0 \leq \tau_1 \leq \tau_2 \leq \dots \leq \tau_k \leq c_\alpha) \\ &= k! \int_0^{c_\alpha} \int_{\tau_1}^{c_\alpha} \dots \int_{\tau_{k-1}}^{c_\alpha} \prod_{i=1}^k g(\tau_i) d\tau_k \dots d\tau_1 \\ &= G^k(c_\alpha) \end{aligned} \tag{31}$$

where  $G(c_\alpha)$  is the (cumulative) distribution function of the  $F_{T-k}^1$  distribution. When  $k = 40$  with 140 degrees of freedom, (31) yields:

$$P(\tau_i \leq 3.59, i = 1, \dots, 40) = G^{40}(3.59) = (1 - 0.06)^{40} = 0.080 \tag{32}$$

matching (30).

However, since there will be many ‘highly insignificant’ variables in a set of 40 irrelevant regressors, the bound of  $\widehat{t}_{(i)}^2 < 4.99$  is probably the binding one, yielding (at the ‘average’ of 120 degrees of freedom),  $P(t_{(i)}^2 < 4.99 \forall i) \simeq 0.3$ . Reducing both  $T$  and  $k$  need not in fact improve the chances of correct selection when  $k/T$  rises: for example,  $T = 80$ ,  $c = 1$  and  $k = 30$  leads to a range between  $P(t_{(i)}^2 \leq 2.82, \forall i = 1, \dots, 30) \simeq 0.04$  and  $P(t_{(i)}^2 \leq 4.45, \forall i) \simeq 0.31$ . 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.

The first is reducing the maximum size of model to be considered using ‘pre-selection’ as in (say) Hansen (1999). He enforces a maximum of 10 variables in the *SIC* formula when  $T = 140$ , despite  $n = 40$  initially, by sequentially eliminating variables with the smallest  $t$ -values until 30 are removed. However, such a procedure entails that *SIC* actually confronts a different problem, namely a penalty function applied as if  $n = 10$ , so we consider the consequences of that step. If pre-selection did not matter, then under the null, when  $k = 10$  with 130 degrees of freedom, (29 delivers  $\widehat{t}_{(10)}^2 \leq 4.67$  so:

$$P(\tau_i \leq 4.67, i = 1, \dots, 10) = G^{10}(4.67) = (0.9675)^{10} = 0.72. \tag{33}$$

Using the ‘baseline’  $F$ -value of 3.59 (from  $n = 40$ ) in (30) yields:

$$P(t_{(i)}^2 \leq 3.59, \forall i = 1, \dots, 10) = 0.54 \tag{34}$$

so even allowing for the initial existence of 40 variables matters considerably. But the retained variables are those selected to have the largest  $t$ -values out of the whole set of  $k = 40$  (not just  $k = 10$ ), so (33)

overstates the likely performance of Hansen's approach. Conversely (30) will understate what happens after pre-selection, because the very act of altering  $n$  changes the parameters of  $SIC$ , and is not just a 'numerical implementation'. Hansen (his Table 1) reports a probability of 0.45 for correctly locating the null model when  $c = 1$  in his Monte Carlo applied to Hoover-Perez experiments, when excluding the 30 variables with the smallest t-values irrespective of their significance.

Formally, the conditional probability of the 10 largest squared t-values being insignificant at the critical value  $c_\alpha^*$  entailed by (29) applied as if  $k = 10$ , given that the  $k_1 = 30$  smallest squared t-values have been excluded as insignificant is:

$$\begin{aligned} & P(\tau_i \leq c_\alpha^*, i = k_1 + 1, \dots, k \mid \tau_i \leq c_\alpha^*, i = 1, \dots, k_1) \\ &= \frac{P(0 \leq \tau_1 \leq \tau_2 \leq \dots \leq \tau_k \leq c_\alpha^*)}{P(0 \leq \tau_1 \leq \tau_2 \leq \dots \leq \tau_{k_1} \leq c_\alpha^*)}. \end{aligned} \quad (35)$$

The numerator is (31), and the denominator is obtained from the marginal density of  $\tau_1, \dots, \tau_{k_1}$  which is:

$$g_{\tau_1, \tau_2, \dots, \tau_{k_1}}(\tau_1, \tau_2, \dots, \tau_{k_1}) = \frac{k!}{(k - k_1)!} (1 - G(\tau_{k_1}))^{k - k_1} \prod_{i=1}^{k_1} g(\tau_i),$$

so:

$$\begin{aligned} & P(0 \leq \tau_1 \leq \tau_2 \leq \dots \leq \tau_{k_1} \leq c_\alpha^*) \\ &= \frac{k!}{(k - k_1)!} \int_0^{c_\alpha} \int_{\tau_1}^{c_\alpha} \dots \int_{\tau_{k_1}}^{c_\alpha} (1 - G(\tau_{k_1}))^{k - k_1} \prod_{i=1}^{k_1} g(\tau_i) d\tau_{k_1} \dots d\tau_1 \\ &= \sum_{i=0}^{k - k_1} \frac{k!}{i! (k - i)!} G^{k - i}(c_\alpha^*) (1 - G(c_\alpha^*))^i. \end{aligned}$$

The probability of excluding the last 10 variables (those with largest t-values) given that we have excluded the first 30 variables, has a denominator:

$$\begin{aligned} & P(0 \leq \tau_1 \leq \tau_2 \leq \dots \leq \tau_{30} \leq c_\alpha^*) \\ &= \sum_{i=0}^{10} \frac{40!}{i! (40 - i)!} G^{40 - i}(4.67) (1 - G(4.67))^i \\ &= \sum_{i=0}^{10} \frac{40!}{i! (40 - i)!} (0.9675)^{40 - i} (0.0325)^i \\ &\simeq 1.0. \end{aligned}$$

Unsurprisingly, the 11<sup>th</sup> largest t-value is not significant under the null, so from (35):

$$\begin{aligned} & P(\tau_i \leq 4.67, i = 31, \dots, 40 \mid \tau_i \leq 4.67, i = 1, \dots, 30) \\ &= \frac{P(0 \leq \tau_1 \leq \tau_2 \leq \dots \leq \tau_k \leq 4.67)}{P(0 \leq \tau_1 \leq \tau_2 \leq \dots \leq \tau_{k_1} \leq 4.67)} \\ &\simeq \frac{0.267}{1} = 0.267, \end{aligned}$$

which is smaller than 0.54 in (34), and smaller than Hansen found by simulation, but much larger than (32). Pre-selecting out the least significant variables improves the performance of  $SIC$  in finite samples,

although it would seem in general preferable to follow the *PcGets* approach of using F-tests to do so, rather than arbitrarily exclude a pre-assigned number of regressors, despite the low probability of their significance under the null. Notice that when  $T = 140$ ,  $k_1 = 0$  and  $k = 10$ , the unconditional value of  $F_{130}^{10}$  from (25) only needs to be less than 5.5 when  $c \geq 1.0$ , which is virtually certain to occur under the null, so block tests have advantages here.

A second approach to raising the chances of correctly selecting the null model is to increase  $c$ . For example,  $c = 2$  raises the required  $\hat{t}_{(i)}^2$  to 7.31 in Hoover–Perez, and hence for  $n = 40$ :

$$P\left(\hat{t}_{(i)}^2 \leq 7.31, \forall i = 1, \dots, 40\right) = (1 - 0.0081)^{40} = 0.73,$$

which is a dramatic improvement over (30). Hansen’s setting of  $c = 2$  when  $n = 10$  further raises the required  $\hat{t}_{(i)}^2$  to 9.51, and ignoring pre-selection would deliver a 97.5% chance of correctly finding a null model. The actual conditional probability for  $c = 2$  and  $c_\alpha^* = 9.51$  is:

$$P(\tau_i \leq 9.51, i = 31, \dots, 40 \mid \tau_i \leq 9.51, i = 1, \dots, 30) = 0.905,$$

which is much larger than for  $c = 1$  (Hansen reports 95% in his Monte Carlo, whereas  $(1 - 0.0081)^{10} = 0.92$ ).

Nevertheless, when the null is false, both steps of raising  $c$  and/or statistically or arbitrarily simplifying till 10 variables remain could greatly reduce the probability of retaining relevant regressors with absolute t-values smaller than 2.5, as Hansen notes. This effect does not show up in his analysis of the Hoover–Perez experiments because the population t-values are either very large or very small. Moreover, there are very few relevant variables in the DGPs of those experiments, whereas  $m > 10$  in (5) would ensure an inconsistent selection.

## 6 Conclusion

The automatic selection algorithms in *PcGets* provide consistent selection rules like *SIC* or *HQ*, based on the logic of the behaviour of information criteria, but mapped to significance levels. However, the asymptotic comfort of consistent selection when the model nests the DGP does not greatly restrict the choice of strategy in small samples. As shown in figures 1–3, there is a very wide range of implicit significance levels across the information criteria, and (e.g.) neither *AIC* nor *HQ* seem well designed for the null-DGP experiments in Hoover and Perez (1999): even *SIC* struggles. In finite samples, *PcGets* both ensures a congruent model and can out-perform in important special cases (such as a null DGP) without *ad hoc* adjustments. Indeed, depending on the state of nature, *PcGets* can have a higher probability of finding the DGP starting from a highly over-parameterized GUM using the Liberal strategy, than commencing from the DGP and selecting by the Conservative strategy (see Hendry and Krolzig, 2003b). 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.

Four other conclusions emerge from this analysis. First, pre-selection can help locate the DGP by altering the ‘parameters’ entered into *SIC* calculations, specifically the apparent degrees of freedom and the implicitly required t-value. *PcGets* employs a statistical ‘pre-selection’ first stage based on block sequential tests, but with loose significance levels, to try and ensure that relevant variables are unlikely

to be eliminated. This greatly improves its performance when the null is true, but also does so more generally. The findings reported in Hansen (1999) for other cases suggest that a similar result holds in finite samples for *SIC*. Secondly, the trade-off between retaining irrelevant and losing relevant variables remains for information criteria, and is determined by the choice of  $c$  in *SIC* implicitly altering the significance level. In problems with many  $t$ -values around 2 or 3, high values of  $c$  will be detrimental, and almost surely not compensated by benefits achieved when the DGP is null. Thirdly, *SIC* does not address the difficulty that the initial model specification may not be adequate to characterize the data, and *SIC* will 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 lowers the relative success rate of *PcGets*. Finally, the arbitrary specification of an upper bound on  $n$  is both counter to the spirit of *SIC*, and would deliver adverse findings in any setting where  $n$  was set lower than the number  $m$  of relevant DGP variables.

We have not explored a strategy coinciding with  $AIC_k^c$  but intend to do so, to allow an option for users interested in asymptotically efficient selection.

## 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 Symposium on Information Theory*, pp. 267–281. Budapest: Akademia Kiado.
- Atkinson, A. C. (1981). Likelihood ratios, posterior odds and information criteria. *Journal of Econometrics*, *16*(1), 15–20.
- 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.
- Doornik, J. A. (1999). *Object-Oriented Matrix Programming using Ox*. London: Timberlake Consultants Press. 3rd edition.
- 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. (2001). *Automatic Econometric Model Selection*. London: Timberlake Consultants Press.
- Hendry, D. F., and Krolzig, H.-M. (2003a). New developments in automatic general-to-specific modelling. In Stigum, B. P. (ed.), *Econometrics and the Philosophy of Economics*, pp. 379–419. Princeton: Princeton University Press.
- Hendry, D. F., and Krolzig, H.-M. (2003b). The properties of automatic Gets modelling. Unpublished paper, Economics Department, Oxford University.

- Hendry, D. F., and Krolzig, H.-M. (2003c). Model selection with more variables than observations. Unpublished paper, Economics Department, Oxford University.
- 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.
- Hurvich, C. M., and Tsai, C.-L. (1989). Regression and time series model selection in small samples. *Biometrika*, **76**, 297–307.
- 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.
- Kullback, S., and Leibler, R. A. (1951). On information and sufficiency. *Annals of Mathematical Statistics*, **22**, 79–86.
- Lovell, M. C. (1983). Data mining. *Review of Economics and Statistics*, **65**, 1–12.
- Phillips, P. C. B. (1994). Bayes models and forecasts of Australian macroeconomic time series. In Hargreaves, C. (ed.), *Non-stationary Time-Series Analyses and Cointegration*. Oxford: Oxford University Press.
- Phillips, P. C. B. (1995). Automated forecasts of Asia-Pacific economic activity. *Asia-Pacific Economic Review*, **1**, 92–102.
- Phillips, P. C. B. (1996). Econometric model determination. *Econometrica*, **64**, 763–812.
- Phillips, P. C. B., and Ploberger, W. (1995). An asymptotic theory of Bayesian inference for time series. *Econometrica*, **63**, 381–412.
- Rissanen, J. (1978). Modeling by shortest data description. *Automatica*, **14**, 465–471.
- Robinson, P. M. (2003). Denis Sargan: Some perspectives. *Econometric Theory*, **19**, 481–494.
- Sargan, J. D. (1975). Asymptotic theory and large models. *International Economic Review*, **16**, 75–91.
- Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, **6**, 461–464.
- Shibata, R. (1980). Asymptotically efficient selection of the order of the model for estimating parameters of a linear process. *Annals of Statistics*, **8**, 147–164.
- Sober, E. (2003). Instrumentalism, parsimony, and the Akaike framework. Unpublished paper, Department of Philosophy, University of Wisconsin, Madison.
- Sugiura, N. (1978). Further analysis of the data by Akaike's information criterion and the finite corrections. *Communications in Statistics*, **A78**, 13–26.
- Theil, H. (1961). *Economic Forecasts and Policy*, 2nd edn. Amsterdam: North-Holland Publishing Company.
- White, H. (1990). A consistent model selection. In Granger, C. W. J. (ed.), *Modelling Economic Series*, pp. 369–383. Oxford: Clarendon Press.