ISSN 1471-0498

DEPARTMENT OF ECONOMICS

DISCUSSION PAPER SERIES

EVALUATING AUTOMATIC MODEL SELECTION

Jennifer L. Castle, Jurgen A. Doornik and David F. Hendry

Number 474
January 2010

Manor Road Building, Oxford OX1 3UQ

Evaluating Automatic Model Selection

Jennifer L. Castle, Jurgen A. Doornik and David F. Hendry*
Department of Economics, University of Oxford, UK

Abstract

We evaluate automatically selecting the relevant variables in an econometric model from a large candidate set. General-to-specific selection is outlined for a constant model in orthogonal variables, where only one decision is required to select, irrespective of the number of regressors ($N < T$) where T is the sample size, then evaluated in simulation experiments for $N = 1000$. Comparisons with *Autometrics* (Doornik, 2009) show similar properties, but not restricted to orthogonal cases. Monte Carlo experiments examine the roles of post-selection bias corrections and diagnostic testing, and evaluate *Autometrics*' capability in dynamic models by its costs of search versus costs of inference.
JEL classifications: C51, C22.
KEYWORDS: Model Selection, *Autometrics*, post-selection bias correction, costs of search, costs of inference.

1 Introduction

Our contribution concerns model selection in situations that involve specification uncertainty over the choice of which variables, lags, functional forms etc., are relevant and which irrelevant. To successfully determine what matters and how it enters, all important determinants need to be included, since omitting key variables adversely affects the goodness of fit, biases the included factors' effects, and in a world of intercorrelated variables with non-stationarities induced by breaks, leads to non-constant estimated models. In related research, we have considered modeling non-linearity (see Castle and Hendry, 2009), and multiple breaks (see Castle, Doornik and Hendry, 2009, building on Hendry, Johansen and Santos, 2008, and Johansen and Nielsen, 2009), with an empirical application in Hendry and Mizon (2009), so here we consider why model selection might be successful in general, then focus on dynamic specification.

We consider nine possible criteria for evaluating the success, or otherwise, of model selection, and propose three as both operational and relevant to practical empirical modeling. These 3 criteria are then applied to general-to-specific (*Gets*) selection, first for the analytically tractable setting of a constant model in orthogonal variables, an unknown subset of which are relevant to explaining the dependent variable, and the remainder are irrelevant. We demonstrate that only one selection decision is required irrespective of the number of regressors $N < T$, for T observations. This rebuts claims that model selection intrinsically involves repeated testing

*It is a pleasure to contribute to this volume in honor of Svend Hylleberg. DFH has known Svend since DFH presented his first ever conference paper at the 1969 Econometric Society European Meeting in Brussels. Later, with Timo Teräsvirta, DFH was an external examiner for Svend's 1984 doctorate on seasonality. Although that event is a quarter of a century ago, he still remembers the look on Svend's face during the defence on arguing that Svend's proposed methods were quite unacceptable–pause–as their application to Beethoven's 5th symphony would ruin it! Since then, eight of Svend's top ten citations are to his publications on seasonality. Svend has in turn been an assiduous, and highly successful, doctoral supervisor, with a number of his past students already well-known in their own right. It is rare to find someone who excels in both areas, academia and administration, but, Svend has also had a truly successful career as an administrator, to which arena he has devoted great care and attention, helping to build Aarhus Econometrics into a global force.

(see e.g. Leamer, 1983). A simulation experiment for $N = 1000$ when $T = 2000$ confirms the theoretical analysis underlying what we call the '1-cut' approach. Thus, although there are $2^{1000} \simeq 10^{301}$ possible models, only *one* model needs estimated and only a *single decision* is required to select the final model: 'repeated testing' does not occur.

Because the 'size' of a test statistic has a definition which is only precise for a similar test, and the word is ambiguous in many settings (such as sample size), we use the term '*gauge*' to denote the retention frequency of irrelevant variables when selecting. Similarly, retaining relevant variables by rejecting their null no longer corresponds to the conventional notion of 'power', so we use '*potency*' to denote the average retention frequency of relevant variables. Then, for 1-cut, gauge is close to its corresponding nominal significance level, α, for small α (e.g., $\alpha \leq 1/N$), which can be controlled, and potencies are close to the theoretical powers for one-off tests, despite selecting (e.g.) $n = 10$ relevant variables from $N = 1000$ candidates.

Although there is no repeated testing, *selection* affects the distributional properties of the final model's estimates as compared with estimating the local data generating process (LDGP–the DGP in the space of the variables under analysis: see Hendry, 2009). Thus, the second step is to correct for biases induced in conditional distributions by only retaining significant coefficients. Building on Hendry and Krolzig (2005), we show that on balance, bias corrections improve both the conditional and unconditional distributions of irrelevant variables as measured by mean-square errors (MSEs), with a small increase in the MSEs of relevant variables.

Next, to check the closeness of their results, the 1-cut approach is compared with using a general search algorithm implementing automatic *Gets*, which does not depend on orthogonality of regressors, here *Autometrics* within *PcGive* (see Doornik, 200b7, 2009, Hendry and Doornik, 2009). We also assess the impact of mis-specification testing on gauges, and the role of tests for encompassing the initial general unrestricted model (GUM). We use a much smaller N (namely 10) in these simulation experiments, so results can be graphed and compared across a range of experiments as the number of relevant variables, $n \leq N$, and their significance changes, using the general design from Castle, Qin and Reed (2009).

Since dynamic dependence induces autocorrelations between adjacent lags, the resulting non-orthogonality requires a general algorithm. We extend the experimental design to stationary dynamic DGPs, including models with under-specification, matching the LDGP, and over-specification. Negative dependence (e.g., the levels representation of first differences) can be problematic for selection approaches that do not use *Gets*, such as stepwise expanding searches and the Lasso (see Tibshirani, 1996). *Autometrics* undertakes a lag-length pre-search at loose significance levels from the longest lag, which has little impact on the selection results, but greatly improves the search time for large N. Dynamics *per se* do not seem to affect the selection procedure, but measurements of performance must account for the difficulty in precisely dating lag reactions.

The structure of the paper is as follows. Section 2 considers how to evaluate model selection approaches. Section 3 discusses the 1-cut approach and presents simulation findings when there are $N = 1000 < T = 2000$ orthogonal regressors. Section 4 compares *Autometrics* to 1-cut in Monte Carlos for $N = 10$ across a range of static experiments, followed by an evaluation of the impact of diagnostic and encompassing testing. Section 5 generalizes to the dynamic case for up to 14 regressors and evaluates the costs of inference and search. Section 6 concludes.

2 Evaluating model selection

The properties of empirical models are determined by how they are formulated, selected, estimated, and evaluated, as well as by data quality, the initial subject-matter theory and institutional and historical knowledge. Many features of models are not derivable from subject-matter theory, and in practice empirical evidence is essential to determine what are the relevant vari-

ables, lag reactions, parameter shifts, non-linear functions and so on. All steps are prone to difficulties, even for experts, which is why automatic methods merit consideration. 'Model uncertainty' comprises much more than whether the 'correct model' was selected from some set of candidate variables that nested the LDGP, which essentially assumes the 'axiom of correct specification' for the proposed model. A key aim of model selection is to reduce some of the uncertainties about the many aspects involved in specification, at the cost of a 'local increase' in uncertainty as to precisely which influences should be included and which excluded around the margin of significance. Thus, embedding any claimed theory in a general specification that is congruent with all the available evidence offers a chance to both utilize the best available theory insights and learn from the empirical evidence. However, such embedding can increase the initial model size to a scale where a human has intellectual difficulty handling the required reductions, and indeed the general model may not even be estimable, so computerized, or automatic, methods for model selection become essential.

Nevertheless, the best model selection approaches cannot be expected to select the LDGP on every occasion, even when *Gets* is directly applicable and the GUM nests the LDGP. Conversely, no approach will work well when the LDGP is not a nested special case of the postulated model, especially in processes subject to breaks that induce multiple sources of non-stationarity. Phillips (2003) provides an insightful analysis of the limits of econometrics.

Models that are constructed with a specific purpose in mind need to be evaluated accordingly. Thus, there are many grounds on which to select empirical models—theoretical, empirical, aesthetic, and philosophical—and within each category, many criteria, leading to numerous ways to judge the 'success' of selection algorithms, including:

(A) maximizing the goodness of fit;

(B) recovering the LDGP with high frequency;

(C) improving inference about parameters of interest over the GUM;

(D) improving forecasting over the GUM (and other selection methods);

(E) working well for 'realistic' LDGPs;

(F) matching a theory-derived specification;

(G) recovering the LDGP starting from the GUM almost as often as from the LDGP itself;

(H) matching the operating characteristics of the algorithm with their desired properties;

(I) finding a well-specified, undominated model of the LDGP.

The first is a traditional criterion, often based on penalized fit, but Lovell (1983) showed that it did not lead to useful selections. The second is overly demanding, as it may be nearly impossible to find the LDGP even when commencing from it (e.g., some relevant variables may have $|\mathrm{t}| < 0.1$). The third seeks (e.g.) small, accurate, uncertainty regions around estimated parameters of interest, and has been criticized by Leeb and Pötscher (2003, 2005) among others. There are many contending approaches when (D) is the objective, including using other selection methods, averages over a class of models, factor methods, robust devices, or neural nets. However, in processes subject to breaks, in-sample performance need not be a reliable guide to later forecasting success (see Clements and Hendry, 1998, 1999). There are also many possible contenders for (E), including, but not restricted to, Phillips (1994, 1995, 1996), Tibshirani (1996), Hoover and Perez (1999, 2004), Hendry and Krolzig (1999, 2001), White (2000), Krolzig (2003), Kurcewicz and Mycielski (2003), Demiralp and Hoover (2003), and Perez-Amaral, Gallo and White (2003), as well as stepwise regression, albeit most with different properties in different states of nature. Criterion (F) is again widely used, and must work well

when the LDGP coincides with the theory model, but otherwise need not. For (G), a distinction must be made between costs of inference and costs of search. The former apply to commencing from the LDGP, so confront even an investigator who did so, but who was uncertain that the specification was completely correct (omniscience is not realistic in empirical economics), and are inevitable when test rejection frequencies are non-zero under the null, and not unity for all alternatives. Costs of search are additional, due to commencing from a GUM that nests but is larger than the LDGP, so are really due to selecting. Operating characteristics for (H) could include that the nominal null rejection frequency matches the gauge; that retained parameters of interest are unbiasedly estimated; that MSEs are small, etc. Finally, there is the 'internal criterion' (I) that the algorithm could not do better for the given sample, in that no other model dominates that selected. We use (G), (H) and (I) as the basis for evaluation, noting that they could in principle be achieved together.

3 Why *Gets* model selection can succeed

When all the regressors are mutually orthogonal, it is easy to explain why *Gets* model selection needs only a single decision. Consider the perfectly orthogonal regression model:

$$y_t = \sum_{k=1}^{N} \beta_k x_{k,t} + \epsilon_t \tag{1}$$

where $T^{-1}\sum_{t=1}^{T} x_{k,t}x_{j,t} = \lambda_k \delta_{k,j}\ \forall k, j$, where $\delta_{k,j} = 1$ if $k = j$ and zero otherwise, with $\epsilon_t \sim \mathsf{IN}[0, \sigma_\epsilon^2]$, independently of the $\{x_{k,t}\}$, and $T >> N$. After unrestricted estimation of (1), order the N sample t^2-statistics testing H_0: $\beta_k = 0$ as:

$$\mathsf{t}^2_{(1)} \geq \mathsf{t}^2_{(2)} \geq \cdots \geq \mathsf{t}^2_{(N)} \tag{2}$$

The cut-off, $\widetilde{n}$, between retained and excluded variables using a 2-sided significance level c_α for a t-test is:

$$\mathsf{t}^2_{(\widetilde{n})} \geq c^2_\alpha > \mathsf{t}^2_{(\widetilde{n}+1)}. \tag{3}$$

Variables with large t^2 values are retained and all other variables are eliminated. Only *a single decision* is needed to implement (3), even for $N = 1000$, and 'repeated testing' does not occur. Using this 1-cut decision rule, it is straightforward to maintain the false null retention rate at (say) less than one variable by setting $\alpha \leq 1/N$, $\forall N$ (for small N, much tighter choices are feasible): α should also tend to zero as T increases to ensure a consistent selection (see Hannan and Quinn, 1979, Pötscher, 1991, and Campos, Hendry and Krolzig, 2003).

In non-orthogonal problems, path search is required to establish 'genuine relevance', which gives the impression of 'repeated testing', but should not be confused with selecting the 'best fitting model' from the $2^{1000} \simeq 10^{301}$ possible models. *Autometrics* uses a tree-path search to detect and eliminate statistically-insignificant variables, thereby improving on the multi-path procedures in Hoover and Perez (1999) or Hendry and Krolzig (2001). Such an algorithm does not become stuck in a single-path sequence, where a relevant variable is inadvertently eliminated, retaining other variables as proxies (e.g., as in stepwise regression). At any stage, a variable removal is only accepted if the new model is a valid reduction of the GUM (i.e., the new model must encompass the GUM at the chosen significance level: see Doornik, 2008). A path terminates when no variable meets the reduction criterion. At the end, there will be one or more non-rejected (terminal) models: all are congruent, undominated, mutually-encompassing representations. If necessary, the search is terminated using a tie-breaker, e.g., the Schwarz (1978) information criterion, although all terminal models are reported and can be used in,

say, forecast combinations. Thus, goodness-of-fit is not directly used to select models, and no attempt is made to 'prove' that a given set of variables matters although the choice of c_α affects R^2 and $\widetilde{n}$ through retention by $\mathsf{t}^2_{(\widetilde{n})} \geq c^2_\alpha$. Generalizations are feasible to instrumental variables estimators (see Hendry and Krolzig, 2005), and likelihood estimation (Doornik, 2009).

3.1 Empirical evaluation of model selection

Let the first n regressors be relevant, with $N-n$ irrelevant regressors in the GUM, and let $\widetilde{\beta}_{k,i}$ denote the OLS coefficient estimate after selection for the coefficient on $x_{k,i}$ in the GUM for replication i, with M replications. When $1(\cdot)$ is the indicator variable, potency and gauge respectively calculate the retention frequencies of relevant and irrelevant variables as:

$$\begin{array}{lcl} \text{retention rate: } \widetilde{\mathsf{p}}_k & = & \frac{1}{M}\sum_{i=1}^{M} 1(\widetilde{\beta}_{k,i} \neq 0), \quad k = 1,\ldots,N, \\ \text{potency} & = & \frac{1}{n}\sum_{k=1}^{n} \widetilde{\mathsf{p}}_k, \\ \text{gauge} & = & \frac{1}{N-n}\sum_{k=n+1}^{N} \widetilde{\mathsf{p}}_k. \end{array} \tag{4}$$

In addition, we also compute MSEs, both before and after model selection. Define $\mathcal{M}_g$ as the model obtained after selection from the GUM and $\mathcal{M}_d$ as the model retained after selection from the LDGP. The unconditional and conditional MSEs respectively are calculated as:

$$\begin{aligned} \mathsf{UMSE}_k &= \frac{1}{M}\sum_{i=1}^{M}\left(\beta^*_{k,i} - \beta_k\right)^2, \ \forall k \\ \mathsf{CMSE}_k &= \frac{\sum_{i=1}^{M}\left[\left(\beta^*_{k,i} - \beta_k\right)^2 \cdot 1(\beta^*_{k,i} \neq 0)\right]}{\sum_{i=1}^{M} 1(\beta^*_{k,i} \neq 0)}, \quad (\beta^2_{k,i} \text{ when } \sum_{i=1}^{M} 1(\beta^*_{k,i} \neq 0) = 0) \end{aligned}$$

where $\beta^*_{k,i}$ denotes the coefficient defined in Table 1.

	Coefficient	MSE	Note
GUM	$\widehat{\beta}_{k,i}$	GMSE_k	for $N < T$
$\mathcal{M}_g$	$\widetilde{\beta}_{k,i}$	USMSE_k, CSMSE_k	$\widetilde{\beta}_{k,i} = 0$ if x_k not selected
DGP	$\overline{\beta}_{k,i}$	LMSE_k	$\overline{\beta}_{k,i} = 0$ for $k = n+1, ..., N$
$\mathcal{M}_d$	$\overline{\overline{\beta}}_{k,i}$	UIMSE_k, CIMSE_k	$\overline{\overline{\beta}}_{k,i} = 0$ if x_k not selected or $k > n$

Table 1: MSEs before and after model selection

The square roots of the MSEs are denoted RMSEs. When the GUM nests the LDGP, the difference between $\mathcal{M}_g$ and $\mathcal{M}_d$ is a measure of over-specification. When the GUM does not nest the LDGP (under-specification), the difference between $\mathcal{M}_g$ and $\mathcal{M}_d$ is a measure of mis-specification. Section 5.2 relates the costs of search and inference to Table 1.

3.2 Selection effects and bias corrections

The estimates from the selected model do not have the same properties as if the LDGP equation had simply been estimated. Conditional estimates of relevant variables' coefficients are biased away from the origin as they are only retained when $\mathsf{t}^2 \geq c^2_\alpha$, and irrelevant variables will have $\mathsf{t}^2 \geq c^2_\alpha$ with probability $\alpha\,(N-n)$ (adventitiously significant). Bias correction is straightforward (see Hendry and Krolzig, 2005), which also drives irrelevant variables' coefficients towards the origin, reducing their MSEs. However, sampling also entails that some relevant variables will by chance have $\mathsf{t}^2 < c^2_\alpha$ in the given sample, so not be selected.

Let $\sigma^2_{\widehat{\beta}} = \mathsf{E}[\widehat{\sigma}^2_{\widehat{\beta}}]$ be the population standard error for the OLS estimator $\widehat{\beta}$, and approximate:

$$\mathsf{t}_{\widehat{\beta}} = \frac{\widehat{\beta}}{\widehat{\sigma}_{\widehat{\beta}}} \simeq \frac{\widehat{\beta}}{\sigma_{\widehat{\beta}}} \sim \mathsf{N}\left[\frac{\beta}{\sigma_{\widehat{\beta}}}, 1\right] = \mathsf{N}\left[\psi, 1\right]$$

where $\psi = \beta/\sigma_{\widehat{\beta}}$ is the non-centrality parameter of the t-test. Let $\phi(x)$ and $\Phi(x)$ denote the normal density and its integral, then the expectation of the truncated t-value for a post-selection estimator $\widetilde{\beta}$ such that $|\mathsf{t}_{\widetilde{\beta}}| > c_\alpha$ is (see e.g., Johnson and Kotz, 1970, ch. 13):

$$\psi^* = \mathsf{E}\left[\mathsf{t}_{\widetilde{\beta}} \mid \left|\mathsf{t}_{\widetilde{\beta}}\right| > c_\alpha; \psi\right] = \psi + \frac{\phi(c_\alpha - \psi) - \phi(-c_\alpha - \psi)}{1 - \Phi(c_\alpha - \psi) + \Phi(-c_\alpha - \psi)} = \psi + r(\psi, c_\alpha) \tag{5}$$

Then, (e.g.) for $\psi > 0$:

$$\mathsf{E}\left[\widetilde{\beta} \mid \widetilde{\beta} \geq \sigma_{\widetilde{\beta}} c_\alpha\right] = \beta + \sigma_{\widetilde{\beta}} r(\psi, c_\alpha) = \beta\left(1 + \psi^{-1} r(\psi, c_\alpha)\right) \tag{6}$$

so an unbiased estimator after selection is:

$$\widetilde{\widetilde{\beta}} = \widetilde{\beta}\left(\frac{\psi}{\psi + r(\psi, c_\alpha)}\right) = \widetilde{\beta}\left(\frac{\psi}{\psi^*}\right). \tag{7}$$

Implementation requires an estimate $\widetilde{\psi}$ of ψ based on estimating ψ^* from the observed $\mathsf{t}_{\widetilde{\beta}}$ and solving iteratively for ψ from (5), as in Hendry and Krolzig (2005):

$$\psi = \psi^* - r(\psi, c_\alpha) \tag{8}$$

First replace $r(\psi, c_\alpha)$ in (8) by $r(\mathsf{t}_{\widetilde{\beta}}, c_\alpha)$, and ψ^* by $\mathsf{t}_{\widetilde{\beta}}$:

$$\widetilde{\mathsf{t}}_{\widetilde{\beta}} = \mathsf{t}_{\widetilde{\beta}} - r\left(\mathsf{t}_{\widetilde{\beta}}, c_\alpha\right), \text{ then } \widetilde{\widetilde{\mathsf{t}}}_{\widetilde{\beta}} = \mathsf{t}_{\widetilde{\beta}} - r\left(\widetilde{\mathsf{t}}_{\widetilde{\beta}}, c_\alpha\right) \tag{9}$$

leading to the bias-corrected parameter estimate:

$$\widetilde{\widetilde{\beta}} = \widetilde{\beta}\left(\widetilde{\widetilde{\mathsf{t}}}_{\widetilde{\beta}} / \mathsf{t}_{\widetilde{\beta}}\right). \tag{10}$$

Hendry and Krolzig (2005) show that most of the selection bias is corrected for relevant retained variables by (10), at the cost of a small increase in their conditional MSEs. Thus, correction exacerbates the downward bias in the unconditional estimates of the relevant coefficients, and also increases their MSEs somewhat. Against such costs, bias correction considerably reduces the MSEs of the coefficients of any retained irrelevant variables, giving a substantive benefit in both their unconditional and conditional distributions. Thus, despite selecting from a large set of potential variables, nearly unbiased estimates of coefficients can be obtained with little loss of efficiency from testing irrelevant variables, but suffering some loss from not retaining relevant variables at large values of c_α. As the normal distribution has 'thin tails', the power loss from tighter significance levels is usually not substantial, and although it could be for fat-tailed error processes, Castle *et al.* (2009) show that impulse-indicator saturation (see Hendry *et al.*, 2008, and Johansen and Nielsen, 2009) is a successful antidote.

3.3 Monte Carlo simulation for $N = 1000$

We illustrate the above theory by simulating 1-cut selection from 1000 variables. The DGP is:

$$\begin{aligned} y_t &= \beta_1 x_{1,t} + \cdots + \beta_{10} x_{10,t} + \epsilon_t & (11)\\ \mathbf{x}_t &\sim \mathsf{IN}_{1000}\left[\mathbf{0}, \mathbf{I}\right] & (12)\\ \epsilon_t &\sim \mathsf{IN}\left[0, 1\right] & (13) \end{aligned}$$

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
β	0.063	0.079	0.095	0.111	0.126	0.142	0.158	0.174	0.190	0.206
ψ	2	2.5	3	3.5	4	4.5	5	5.5	6	6.5
$\mathsf{P}_{0.01}$	0.281	0.468	0.662	0.821	0.922	0.973	0.992	0.998	1.000	1.000
$\mathsf{P}_{0.001}$	0.097	0.212	0.382	0.579	0.758	0.885	0.955	0.986	0.997	0.999

Table 2: Coefficients β_k, non-centralities ψ_k, and theoretical retention probabilities.

where $\mathbf{x}_t' = (x_{1,t}, \cdots, x_{1000,t})$. The regressors are only orthogonal in expectation, but are kept fixed between experiments with $T = 2000$. The DGP coefficients and non-centralities, ψ, are reported in Table 2, together with the theoretical powers of t-tests on the individual coefficients.

The GUM, which is the starting point for model selection, consists of all 1000 regressors and an intercept (which is also irrelevant here):

$$y_t = \beta_0 + \beta_1 x_{1,t} + \cdots + \beta_{1000} x_{1000,t} + u_t, \quad t = 1, \ldots, 2000.$$

Only the first $n = 10$ variables relevant, so 991 variables are irrelevant in the GUM. Selection is undertaken by ordering the t^2s as in (2), retaining (discarding) all variables with t^2-statistics above (below) the critical value as in (3), so selection is made in one decision. We report the outcomes for $\alpha = 1\%$ and 0.1% using $M = 1000$ replications.

Gauges and potencies are recorded in Table 3. Gauges are not significantly different from their nominal sizes, α, so selection is correctly 'sized', and potencies do not deviate from the average powers of 0.81 and 0.69. Thus, there is a close match between theory and evidence, even when selecting 10 relevant regressors from 1000 candidate variables.

α	Gauge	Potency
1%	1.01%	81%
0.1%	0.10%	69%

Table 3: Potency and gauge for 1-cut selection with 1000 variables.

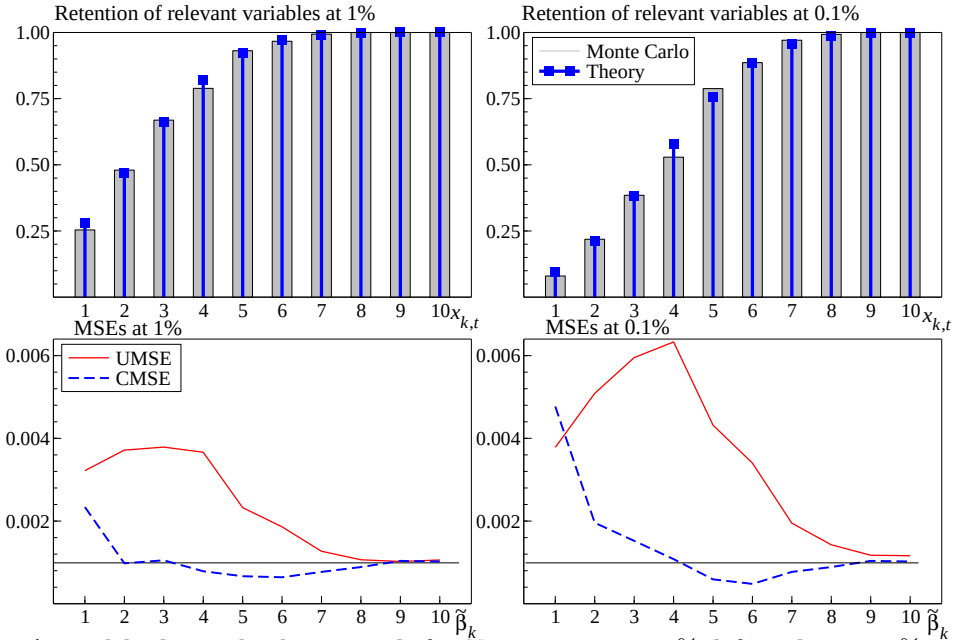

Figure 1: Model selection by the 1-cut rule for $N = 1000$ at $\alpha = 1\%$ (left) and $\alpha = 0.1\%$ (right): retention rates $\widetilde{p}_k$ of relevant variables $x_1, \ldots, x_{10}$ (top graphs), USMSE_k and CSMSE_k (bottom graphs).

Figure 1 records the retention rates of relevant variables against the theoretical retention probabilities: retention rates for individual relevant variables are close to the theoretical powers of individual t-tests, despite selecting from 10^{301} possible models. The CSMSEs are always below the USMSEs for the relevant variables (bottom graphs in Fig. 1), with the exception of β_1 at 0.1%. Baseline USMSEs for all estimated coefficients in the GUM are 0.001 as shown.

3.4 Impact of bias correction on MSEs

In 1-cut selection, all retained variables must be significant at c_α. However, with automated *Gets*, this is not necessarily the case: irrelevant variables may be retained because of diagnostic tracking (i.e., a variable is insignificant, but its deletion would make a diagnostic test significant), or because of encompassing (a variable can be individually insignificant, but not jointly with all variables deleted so far). As retained variables with |t|-values less than c_α are in a sense irrelevant, and the bias correction formula in (10) is non-linear at c_α, we apply it only to significant retained variables, setting insignificant variables' coefficients to zero.

α	1%	0.1%	1%	0.1%
	average CSMSE over 990 irrelevant variables		average CSMSE over 10 relevant variables	
uncorrected $\widetilde{\beta}$	0.84%	1.23%	0.10%	0.14%
$\widetilde{\widetilde{\beta}}$ after bias correction	0.38%	0.60%	0.12%	0.13%

Table 4: Average CSMSE of selected relevant and irrelevant variables (excluding β_0), with and without bias correction, $M = 1000$.

Table 4 shows that the bias corrections for the retained irrelevant variables substantially reduce their CSMSEs by downweighting chance significance; since 99.9% of irrelevant variables are always eliminated at $\alpha = 0.001$, their USMSEs are negligible. Figure 2 graphs the MSEs of the bias-corrected relevant coefficient estimates in their conditional distributions. Here, the impact of bias correction can also be beneficial, but is generally small.

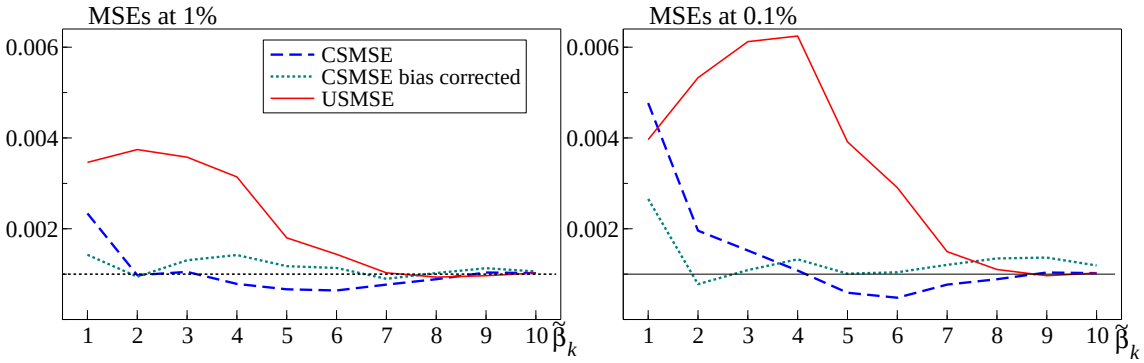

Figure 2: Impact of bias correction on CSMSE_k for relevant variables at $\alpha = 1\%$ (left) and $\alpha = 0.1\%$ (right).

4 1-cut selection and *Autometrics* comparisons

We now compare the 1-cut rule with *Autometrics* using an experiment with a much smaller number of candidate regressors based on a design formulated by Castle *et al.* (2009). This enables visual inspection of the outcomes and covers a wider range of both non-centralities and ratios of relevant to irrelevant variables.

Their experimental design is given by $N = 10$, $n = 1, \ldots, 10$, and $T = 75$:

$$y_t = \beta_0 + \beta_1 x_{1,t} + \cdots + \beta_{10} x_{10,t} + \epsilon_t, \quad (14)$$

$$\mathbf{x}_t \sim \mathsf{IN}_{10}\left[\mathbf{0}, \mathbf{I}_{10}\right], \quad (15)$$

$$\epsilon_t \sim \mathsf{IN}\left[0, \sigma_\psi^2\right], \; t = 1, \ldots, T \quad (16)$$

where $\mathbf{x}_t' = (x_{1,t}, \cdots, x_{10,t})$. The $\mathbf{x}_t$ are fixed across replications as before. Equations (14)–(16) specify 10 different DGPs, indexed by n, each having n relevant variables with $\beta_1 = \cdots = \beta_n = 1$ and $10 - n$ irrelevant variables ($\beta_{n+1} = \cdots = \beta_{10} = 0$). Throughout, we set $\beta_0 = 5$ and $M = 1000$ replications are undertaken. Also, $\sigma_\psi^2 = T/\psi^2$, where ψ denotes the non-centrality, taking the values $2, \ldots, 6$, such that all relevant variables in each experiment have the same non-centrality. The GUM is the same for all 10 DGPs:

$$y_t = \beta_0 + \beta_1 x_{1,t} + \cdots + \beta_{10} x_{10,t} + u_t.$$

Table 5 reports the theoretical powers of t-tests for the ψs considered.

$\alpha \backslash \psi$	2	3	4	5	6
5%	50.3	84.3	97.8	99.9	100
1%	26.0	63.9	91.3	99.1	100

Table 5: Theoretical power for a single t-test (%) for experiment (14)-(16).

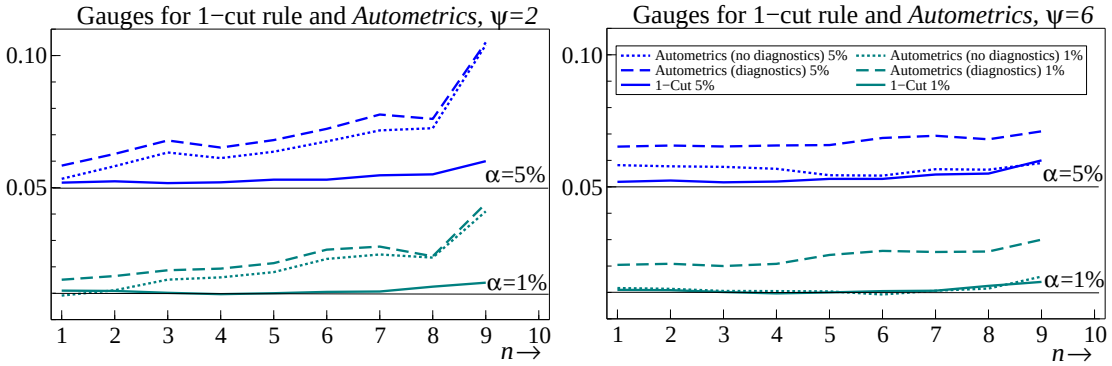

Figure 3: Gauges for 1-cut rule (solid lines), *Autometrics* with diagnostic tracking (dashed lines) and *Autometrics* without diagnostic tracking (dotted lines) for $\alpha = 0.01, 0.05$. The left panel corresponds to $\psi = 2$ and the right panel to $\psi = 6$. The horizontal axis represents the $n = 1, \ldots, 10$ DGPs, each with n relevant variables (and $10 - n$ irrelevant).

We now investigate how the general search algorithm in *Autometrics* performs relative to 1-cut selection in terms of (G)–(I) in Section 2. Their comparative gauges for $\psi = 2$ and $\psi = 6$ are shown in Figure 3 where *Autometrics* selects both without diagnostic tracking and with. In default mode (i.e., with diagnostic tracking), *Autometrics* is 'over-gauged', particularly for low non-centralities, where the gauge increases as $n \to N$. For high non-centralities, the default-mode gauge is increased by about 1-2 percentage points (see Section 4.1). Doornik (2008) shows that encompassing checks against the GUM help stabilize performance.

Figure 4 compares potencies for the 1-cut rule and *Autometrics* without diagnostic tracking. Potency is calculated excluding the intercept which is a 'forced' regressor (i.e., always included) in both algorithms. Hence, potencies can be compared to the power of a single t-test, also recorded in Figure 4 (powers for high non-centralities are excluded as they are close to unity). Both *Autometrics* and the 1-cut rule have potencies close to the optimal single t-test with no selection. *Autometrics* consistently has a higher potency than the 1-cut rule, but potencies are not gauge-corrected and *Autometrics* has a slightly higher gauge. Given this trade-off, there is little difference between the 1-cut rule and searching many paths as in *Autometrics*.

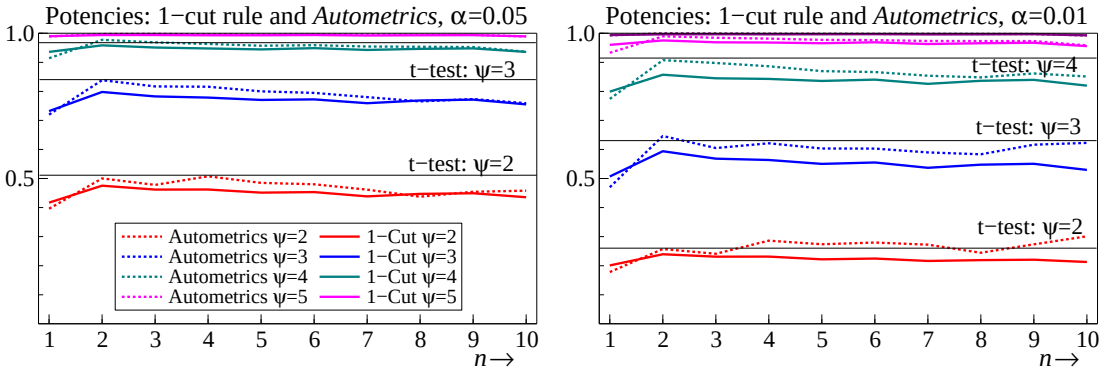

Figure 4: Potencies for 1-cut rule (solid lines) and *Autometrics* without diagnostic tracking (dotted lines) for $\alpha = 0.05$ (left panel) and 0.01 (right panel). The horizontal axis represents the $n = 1, \ldots, 10$ DGPs, each with n relevant variables (and $10 - n$ irrelevant). Solid thin lines record the power for a single t-test at $\psi = 2, 3, 4$.

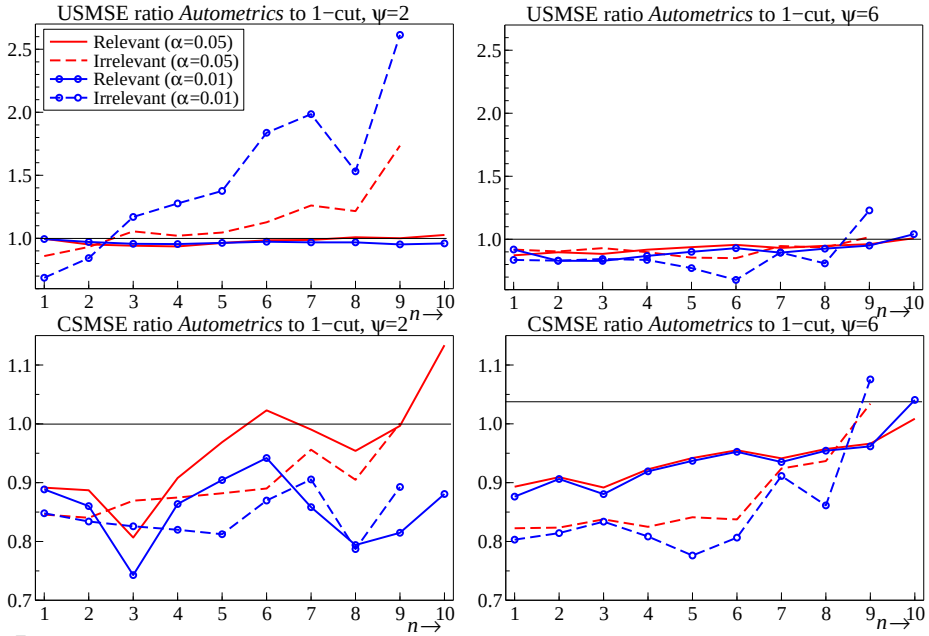

Figure 5: Ratios of MSEs for *Autometrics* to 1-cut rule as n changes, averaging across all relevant (solid lines) and irrelevant (dashed lines) variables. Left-hand panels correspond to $\psi = 2$ and right-hand panels to $\psi = 6$.

Figure 5 records the ratios of MSEs of *Autometrics* selection to the 1-cut rule for both unconditional and conditional distributions, with no diagnostic tests and no bias correction, for $M = 1000$. If the ratio is below unity, *Autometrics* has a smaller average MSE than 1-cut. The lines labelled *Relevant* report the ratios of average MSEs over all relevant variables for a given n. Analogously, the lines labelled *Irrelevant* are based on the average MSEs of the irrelevant variables for each DGP (none when $n = 10$). Unconditionally, the ratios are close to 1 for the relevant variables, but the 1-cut rule performs better for irrelevant variables when the non-centrality is low. When the non-centrality is high, *Autometrics* outperforms the 1-cut rule; the benefits to selection are largest when there are few relevant variables that are highly significant. Conditionally, *Autometrics* outperforms the 1-cut rule in almost all cases–most lines are below unity. There is little loss from using the path-search algorithm even when 1-cut is applicable, and most certainly will be in non-orthogonal problems, when 1-cut would be inappropriate as the initial ranking given by (2) will depend on correlations between variables. The overall 'size' of the selection procedure, $1 - (1 - \alpha)^{N-n}$, can be large, but is uninformative about the success of selection that on average correctly eliminates $(1 - \alpha)(N - n)$ irrelevant variables.

4.1 Impact of diagnostic tests

Figure 3 also compared the gauges for *Autometrics* with diagnostic tracking switched on versus off, both with bias correction. The gauge is close to, but slightly over, the nominal significance level when the diagnostic tests are checked to ensure a congruent reduction. With diagnostic tracking switched off, the gauge is close to the nominal significance level. The difference seems due to irrelevant variables proxying chance departures from the null on one of the five mis-specification tests or the encompassing check, and then being retained despite insignificance.

Figure 6 records the ratio of the USMSEs with diagnostic tests switched off to on in the top panel, and the same for the CSMSEs in the bottom panel, averaging within relevant and irrelevant variables (left-hand panels correspond to $\psi = 2$ and right-hand to $\psi = 6$). Switching the diagnostics off generally improves the USMSEs, but worsens the results conditionally, with the impact coming through the irrelevant variables. Switching the diagnostics off leads to fewer irrelevant regressors being retained overall, improving the USMSEs, but those irrelevant variables that are retained are now more significant than with the diagnostics on. The impact is largest at tight significance levels.

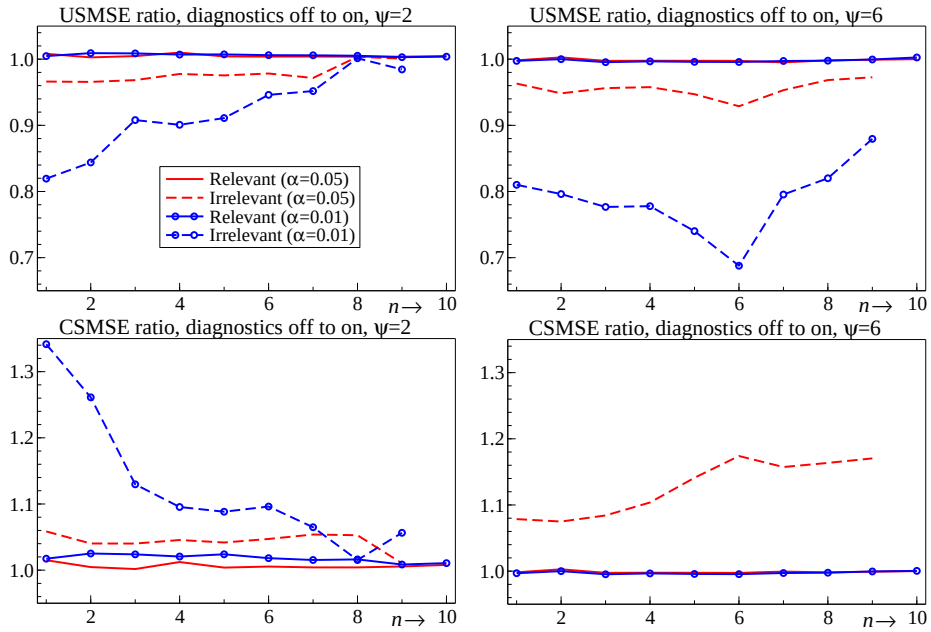

Figure 6: Ratios of MSEs with diagnostic tests off to on for unconditional and conditional distributions as n changes, averaging across all relevant (solid lines) and irrelevant (dashed lines) variables. Left-hand panels correspond to $\psi = 2$ and right-hand panels to $\psi = 6$.

5 Model selection in ADL models

We next consider how *Autometrics* performs in dynamic experiments where 1-cut is an invalid procedure. The experimental design haes 9 DGP specifications given by, for $r = 3, \ldots, 8$:

$$\begin{aligned}
\text{DGP0}: \quad y_t &= \epsilon_t \\
\text{DGP1}: \quad y_t &= 0.75 y_{t-1} + \epsilon_t \\
\text{DGP2}: \quad y_t &= 1.5 y_{t-1} - 0.8 y_{t-2} + \epsilon_t \\
\text{DGPr}: \quad y_t &= 1.5 y_{t-1} - 0.8 y_{t-2} + \sum_{j=1}^{r-2} \left(\beta_j x_{j,t} - \beta_j x_{j,t-1} \right) + \epsilon_t
\end{aligned} \tag{17}$$

where $\epsilon_t \sim \mathsf{IN}[0,1]$ and $\mathbf{x}_t = (x_{1,t}, \ldots, x_{6,t})'$ is generated by:

$$\mathbf{x}_t = \rho \mathbf{x}_{t-1} + \mathbf{v}_t \text{ where } \mathbf{v}_t \sim \mathsf{IN}_6[\mathbf{0}, \mathbf{\Omega}] \tag{18}$$

with $\rho = 0.5$, $\omega_{kk} = 1$, and $\omega_{kj} = 0.5, \forall k \neq j$. There are $n = 0, 1, 2, 4, 6, 8, 10, 12, 14$ relevant regressors. The DGP involves negative relations between pairs of exogenous regressors as the first differences matter. We set $\beta_k = \frac{\psi_k}{\sqrt{T}}$, $\forall k = 1, \ldots, L$, in a given experiment, where $L \leq 6$ is the number of contemporaneous exogenous regressors, and the non-centrality, $\psi_k = 8/\sqrt{2k}$, ranges from 5.5 for DGP3 to just over 2 for DGP8.

There are 7 GUMs, given by $s = 0, 1, 2, 5, 10, 15, 20$:

$$y_t = \mu + \sum_{k=1}^{s} \alpha_k y_{t-k} + \sum_{j=1}^{6} \sum_{k=0}^{s} \gamma_{j,k} x_{j,t-k} + e_t. \tag{19}$$

N is the total number of regressors, with $N = 7, 14, 21, 42, 77, 112, 147$, and $T = 100$, so there are four cases with $N < T/2$, one near T, and two with $N > T$, as well as under-specified examples when $s = 0, 1$, for DGP2–DGP8, and $s = 0$ for DGP1, and over-specified cases. We consider all combinations of DGPs and GUMs, creating 56 experiments in total. Selection uses *Autometrics* at $\alpha = 1\%,\ 0.5\%$, both with and without lag pre-selection with diagnostics switched off (as some models are dynamically mis-specified), for $M = 1000$ replications.

5.1 Potency and gauge

Potency calculated using (4) combines the retention of the lagged dependent variables and the exogenous variables, so we also compute potencies for exogenous variables only by averaging retention rates over the $2L$ relevant exogenous variables. This can be compared with the theoretical power for a t-test on individual coefficients, recorded in Table 6. Potencies for under-specified cases have no precise meaning as relevant variables are omitted, so are not reported.

DGP	3	4	5	6	7	8
ψ	5.66	4.00	3.27	2.83	2.53	2.31
$\mathsf{P}_{0.01}$	0.999	0.915	0.739	0.580	0.462	0.376
$\mathsf{P}_{0.005}$	0.997	0.871	0.654	0.483	0.367	0.287

Table 6: Powers for a single t-test.

Figure 7 records the potencies for every DGP and each GUM specification (defined by the lag length, s, commencing at $s = 2$ to avoid under-specification) for selection with lag pre-search. There is a decline in potency as the non-centrality falls (i.e., the DGP increases), but potency is fairly constant across increasing GUM size (s). There is little impact of extending the GUM when the DGP is autoregressive as the non-centralities of the lagged dependent variables (LDV) are high, so even including 20 lags of y has almost no effect on potency.

The differences between significance levels are fairly small due to the overall potency including the lagged dependent variables which have potencies close to 1. However, comparing the potencies for just exogenous regressors against the powers for a single t-test, the potencies are close to, and in some cases higher than, the corresponding t-test power, despite successive positive and negative coefficients of lagged regressors.

Figure 8 records gauges for each DGP and GUM specification. Gauges should be invariant to the number of regressors and non-centralities so the planes should be flat at the given significance level. For DGP0 with no relevant variables, the gauge is close to the nominal significance level and is somewhat tighter for moderate lag lengths. For the DGPs with just lagged

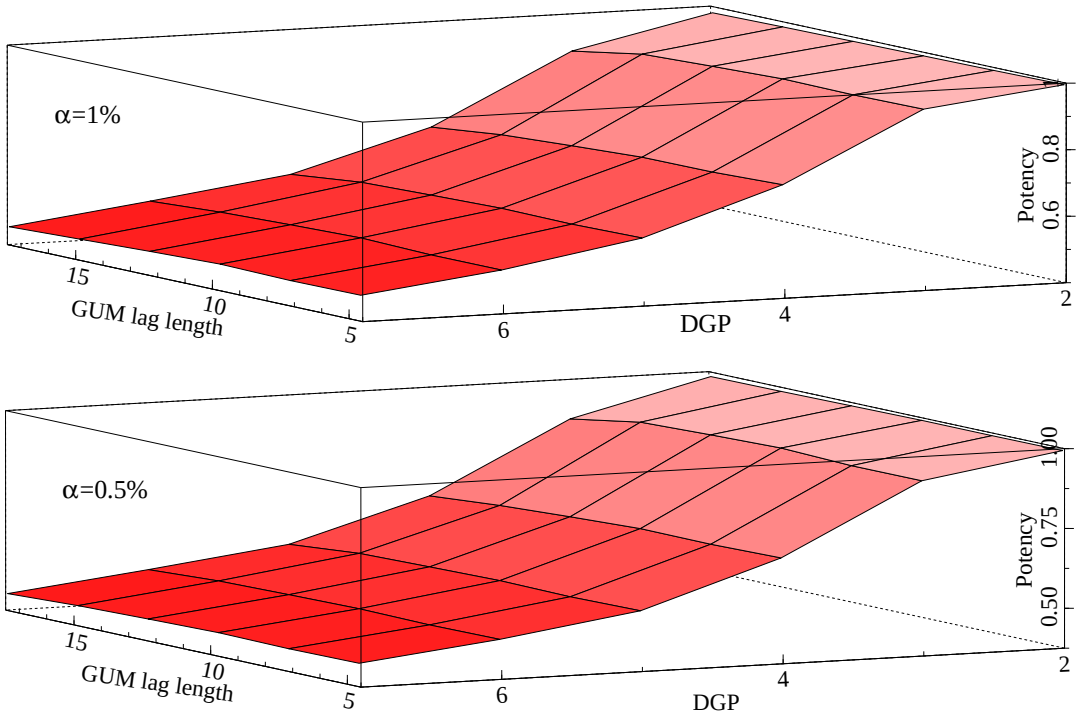

Figure 7: Potency (with pre lag-search) recorded against DGP and GUM specification

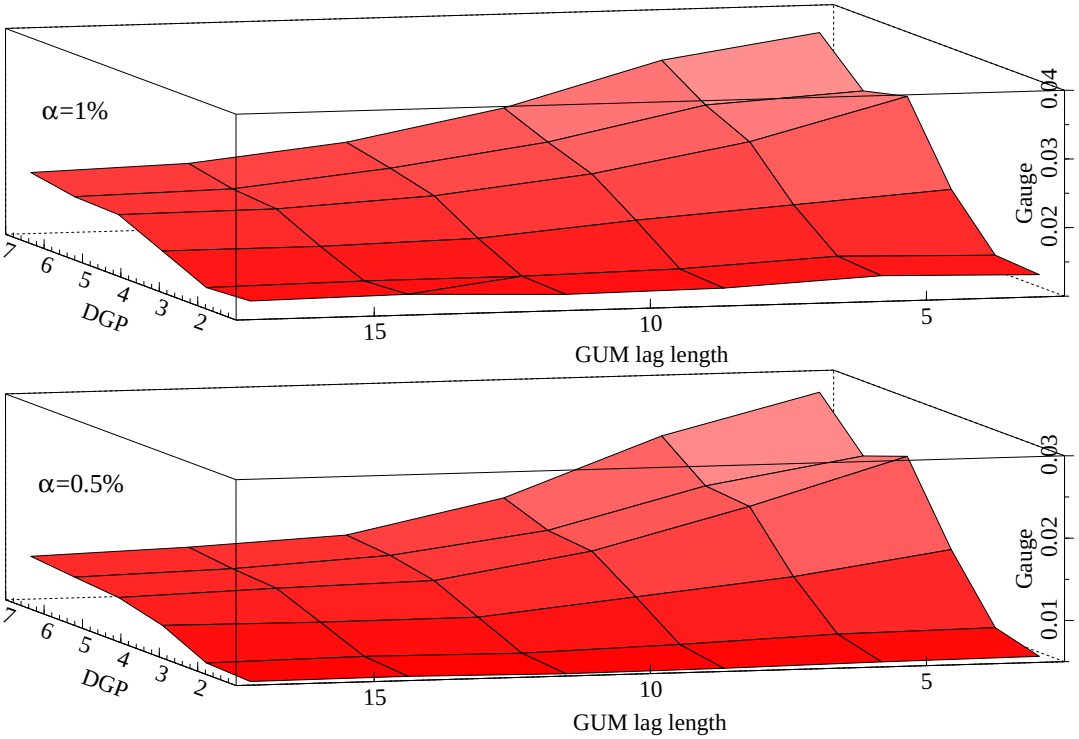

Figure 8: Gauge (with pre lag-search) recorded against DGP and GUM specification

dependent variables (DGP1 and DGP2), the gauges are also close to the nominal significance level, and additional lags do not increase them. The gauges increase as more exogenous regressors become relevant, but the gauges actually fall as the GUM lag length increases. Thus, the gauges are worse for moderate lag lengths ($s = 2$ and 5) than the large GUMs with $s = 15$ or 20. When $s = 15$ or 20 there are more variables than observations so *Autometrics* uses expanding and contracting subsets to reduce the GUM to an estimable intermediate model, see

Doornik (2007a), so despite commencing with $N > T$ the gauge is controlled close to the nominal significance level. Overall, divergences from a flat plane are not substantial.

5.2 Costs of search and inference

We next consider inference and costs of search to assess the selection procedure. We measure costs of inference by the RMSE of applying selection to the LDGP, or conducting inference thereon, namely (see Table 1 for definitions):

$$\sum_{k=1}^{n} \mathsf{UIRMSE}_k. \tag{20}$$

When the GUM is the LDGP, as only significant variables are retained, depending on the choice of critical value, c_α, and the non-centralities, ψ, of the LDGP parameters, (20) could be larger or smaller than the RMSE from direct estimation:

$$\sum_{k=1}^{n} \mathsf{LRMSE}_k. \tag{21}$$

The additional costs of search are calculated as the increase in URMSEs for relevant variables in the selected model when starting from the GUM as against the LDGP, plus the URMSEs computed for all $N - n$ irrelevant variables, both bias corrected:

$$\sum_{k=1}^{n} (\mathsf{USRMSE}_k - \mathsf{UIRMSE}_k) + \sum_{k=n+1}^{N} \mathsf{USRMSE}_k \tag{22}$$

If the LDGP specification is known and just estimated, then $N = n$ and (22) is zero. Otherwise, depending on the non-centralities, ψ, there is usually a trade-off between the two components: as c_α increases, the second falls, but the first may increase. Also, the second rises as $N - n$ increases because it sums over more irrelevant terms. Both seem desirable properties of a measure of search costs. Section 5.4 considers under-specified models, where (22) can be smaller than (20), and may even be negative.

For dynamic models, these measures of search costs evaluate against the precise LDGP lag structure. With substantial autocorrelation, it is difficult to pinpoint the exact timing of relevant lags, so for example y_{t-3} rather than y_{t-2} may be retained. Defining y_{t-3} as an 'irrelevant' variable when y_{t-2} is not retained results in a crude measure of costs. If the selected lags pick up similar dynamics, then search costs would not be as high as indicated by (22). To quantify this, we compute the search and inference costs over the exogenous regressors only, i.e., n becomes $2L$ and N becomes sL where s is the GUM lag length. We separately assess the search costs for the LDVs using:

$$\mathsf{USRMSE}_{\mathsf{LDV}} = \sqrt{\frac{1}{M}\sum_{i=1}^{M}\left(\sum_{k=1}^{s}\widetilde{\beta}_{y,k,i} - \sum_{k=1}^{s}\beta_{y,k,i}\right)^2} \tag{23}$$

where $\widetilde{\beta}_{y,k,i}$ denotes the OLS estimate of the kth lag of the dependent variable. If the retained coefficient estimates sum to the DGP coefficients, then the search costs for the lagged dependent variables would be low despite not selecting the exact lag structure. We compare (23) to the costs of inference, which also sum the retained lagged dependent variables' coefficients when commencing from the LDGP.

Figure 9 records RMSEs for the LDGP given by (21), the costs of inference given by (20) and the costs of search given by (22) over exogenous regressors for each DGP as the GUM

lag length s increases. The costs of search increase as s increases, as there are more irrelevant variables contributing to search costs. These increase steadily (almost linearly for high DGPs) despite a shift from $N \ll T$ to $N > T$ from $s = 10$ to $s = 15$. A tighter significance level results in lower search costs, as fewer irrelevant variables are retained, but delivers higher costs of inference as more relevant variables will be omitted. When there are many irrelevant variables and very few relevant variables that are highly significant (DGP3) the costs of search dominate, but for the larger DGPs (DGP6–DGP8) the costs of search are smaller than the costs of inference for estimable GUMs. Indeed, the costs of search can be smaller than the LDGP costs with no selection (all lower panels). For DGP8 at $\alpha = 0.005$, the costs of search are lower than the costs of inference even for the case where $N > T$ ($s = 15$), so an additional 98 irrelevant variables are searched over. The costs of inference over the LDGP with no selection are quite substantial for the larger DGPs, due to the lower non-centralities.

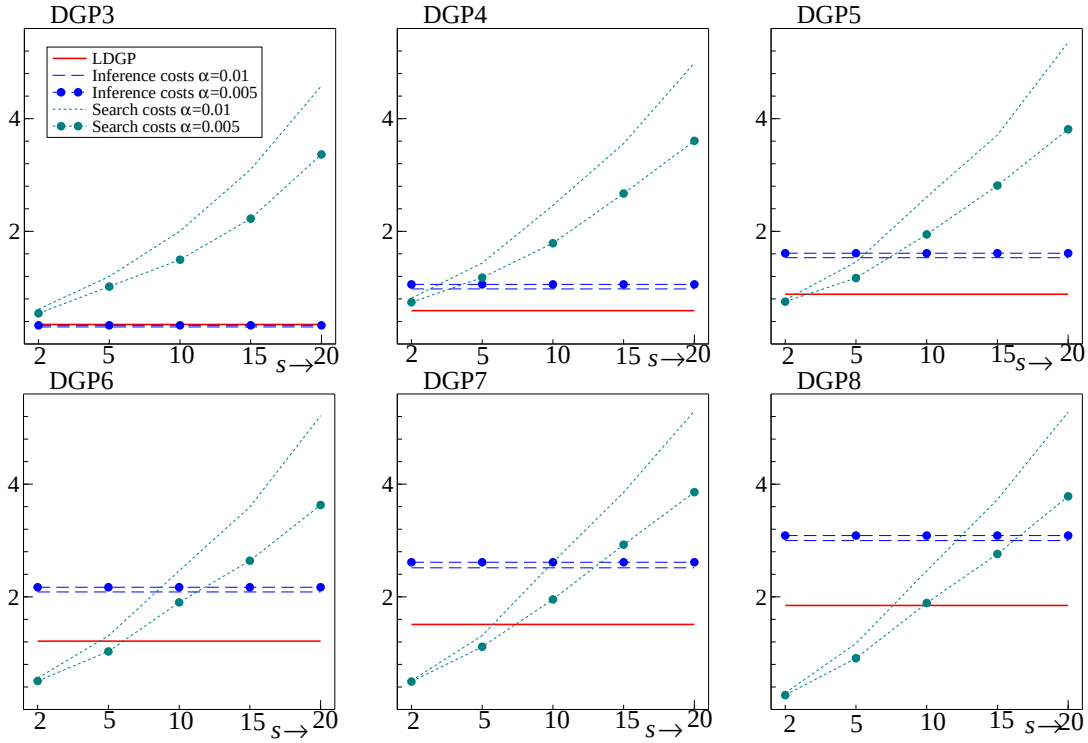

Figure 9: Costs of search and inference for exogenous regressors: LRMSE for the LDGP (solid line), inference on the LDGP (dashed lines) and bias-corrected selection from the GUM with lag pre-selection (dotted lines).

When computing costs for dynamics by averaging over all LDV coefficients, $\mathsf{USRMSE}_{\mathsf{LDV}}$ is close to the equivalent average LRMSE for the LDGP. This is not a pure measure of search costs, but does reflect that the dynamics are adequately captured, although the timing of the dynamics may not be. In practice, timing is likely to only be out by one or at most two lags, depending on data frequency and seasonality. Hence, for dynamic models, selection on average will result in the same long-run solution, but the short-run dynamics may only proxy the LDGP. Thus, the timing of policy impacts, say, will be incorrect but not their overall effect.

5.3 Impact of bias correction and lag pre-search on MSEs

As in the static simulation experiments, we compare the ratios of USMSE and CSMSE with bias correction to without bias correction, averaged over relevant and irrelevant variables. Figure 10 records the average USMSE and CSMSE ratios, averaging across all DGPs, recorded against the GUM specification. All ratios are less than unity, so bias correction is beneficial in all specifications. Most of the benefit comes from down-weighting retained irrelevant vari-

ables, but there is also some advantage to bias correcting the relevant variables. The theory behind these corrections assumes that the only bias source is away from the origin due to selecting just larger t^2 values, whereas inadvertently omitted variables could induce other biases, yet there remains a substantive benefit in practice from bias correction for relevant variables' coefficients, including when $N > T$.

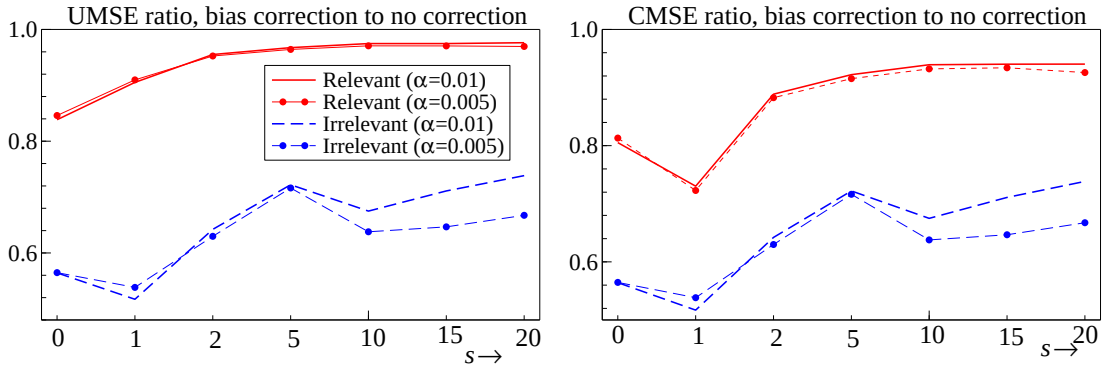

Figure 10: Ratios of MSEs with bias correction to no bias correction, averaged across all relevant (solid lines) and irrelevant (dashed lines) for all DGPs plotted against s.

Lag pre-selection is designed to have no overall impact on the final selected model, and is undertaken at very loose significance levels so as not to eliminate variables that could be relevant when undertaking the tree search. Furthermore, lag pre-search is infeasible when $N > T$ as the initial GUM is inestimable. Computing ratios of USMSEs and CSMSEs with and without lag pre-search results in ratios that are close to unity, although search time is vastly improved with lag pre-selection. There is a small benefit to lag pre-search when the GUM specification includes 5 lags for the irrelevant variables (detailed results available on request).

5.4 Under-specification

For DGP2–DGP8, the GUM is under-specified when $s = 0, 1$, and DGP1 is under-specified when $s = 0$. An LDGP is defined as the joint density of the set of included variables: leaving out any variables that matter defines a different, and obviously less useful, reduction of the DGP. Correlations between variables then lead to included components 'picking up' correlated parts of excluded variables. Evaluating *Autometrics* by how often it finds that under-specified representation would not shed much light on how useful selection would be in practice. Since even the most general formulation is under-specified for the DGP in this section, the equation created by the relevant variables that are included is denoted LDGP* below, but the benchmark for inferences remains the DGP parameters, not the induced parameters of the LDGP.

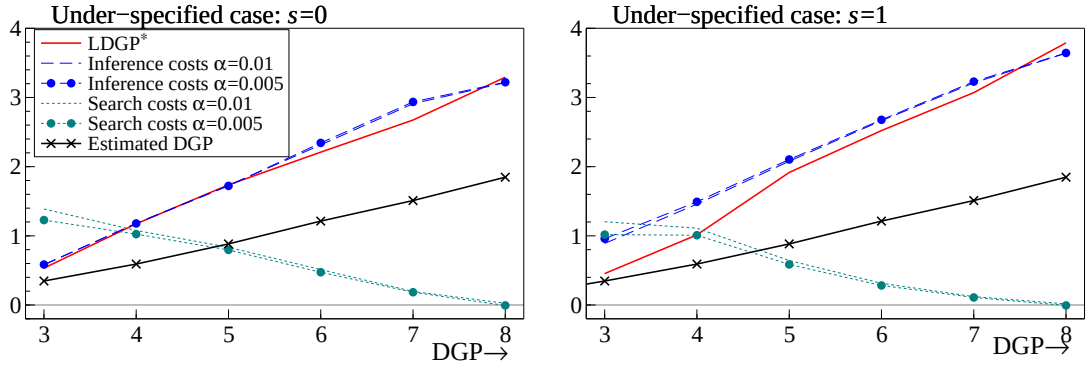

Figure 11: Costs of search and inference for the under-specified case. URMSEs for the LDGP* (solid line), inference on the LDGP* (dashed lines) and bias-corrected selection from the GUM with lag pre-selection (dotted lines).

When models are under-specified for the DGP, the RMSEs for the omitted variables are

their squared DGP parameters, but as these are an additive common element in all models, and in practice it is presumably not known that they are omitted, such terms are excluded in all cost calculations below. Inter-correlations between included and omitted regressors induce biases and inconsistencies in estimated coefficients of the remaining included variables, adding to both search and inference costs. In practice, mis-specification tests may reveal that the LDGP (or GUM) is a poor reduction of the DGP, but that induces a simple-to-general search where it is easy to incorrectly diagnose the source of any rejection (e.g., residual autocorrelation could be due to many mis-specifications), although including the relevant omitted variables would in fact lead to a congruent representation.

Figure 11 records the DGP and LDGP* costs from (21), as well as the costs of inference, (20), and the costs of search, (22), for the LDGP* and GUM in these under-specified cases, again computed only over the exogenous variables and all evaluated against the DGP parameters. As more exogenous variables become relevant, and non-centralities fall, the costs of inference dominate. For $s = 1$ (right-hand panel) the form of mis-specification (the omitted variable is y_{t-2}) is the same for all LDGP*s across the horizontal axis, but the mis-specification has a greater impact as more variables are relevant. In contrast, search costs decline as more variables become relevant (and can even be negative, as in DGP8): the choice of $\alpha = 0.01$ or 0.005 makes almost no difference. The same phenomenon is observed when $s = 0$, with inference costs increasing and search costs decreasing. Thus, there can be higher RMSE costs from just estimating the DGP than from searching in the GUM for the incorrect specification. The GUMs for DGP1 and DGP2 are just a constant for $s = 0$, so are omitted from Figure 11.

5.5 Higher-order dynamics

We extended the simulation exercise to include higher-order dynamics to reflect seasonal dynamics (see Hylleberg, 1986, 1992). In DGP2–DGP8, y_{t-2} is replaced by y_{t-12} to reflect annual lags at the monthly frequency, and the GUM is given by (19) for $s = 15, 20$. A second simulation study replaced y_{t-2} by y_{t-20} for $s = 20$, to reflect 'ice-age' type data measured at 1000-year intervals treating a cycle as 20,000 years. In these cases, lag pre-selection does not help, both because $N > T$ but even if $N \ll T$, the reduction is done on ordered lags. Nevertheless, the gauge, potency and search costs were close to those found for the experiments above, so 'gaps' in the dynamics have little impact: *Autometrics* performs as well on such 'seasonally-dynamic' data as on non-seasonal data.

6 Conclusion

Setting the nominal rejection frequency of individual selection tests at $\alpha \leq 1/N$ where $\alpha \to 0$ as $T \to \infty$, then on average one irrelevant variable will be retained as adventitiously significant out of N candidates. Thus, there is little difficulty in eliminating almost all irrelevant variables when starting from the GUM (a small cost of search). Despite large numbers of irrelevant candidate regressors (including $N > T$), *Autometrics* has a null retention frequency (gauge) close to the nominal size, somewhat increased by undertaking mis-specification testing for congruence and encompassing tests against the GUM. However, bias correction for selection greatly reduces the MSEs of adventitiously retained irrelevant variables in both unconditional and conditional distributions, at a small cost in increased MSEs for relevant variables. The costs of search can be smaller than those of just estimating the DGP even when the GUM is under-specified, and seem to increase only linearly despite $N > T$.

The limits of automatic model selection apply when the LDGP equation would not be reliably selected by the given inference rules applied to itself as the initial specification: selection methods cannot rectify that. Further, when relevant variables have small t-statistics because their parameters are $O(1/\sqrt{T})$, especially when highly correlated with other regressors (see

Leeb and Pötscher, 2003, 2005), then selection is not going to work well: one cannot expect success in selection if a parameter cannot be consistently estimated. Thus, although uniform convergence seems infeasible, selection works for parameters larger than $\mathsf{O}(1/\sqrt{T})$ (as they are consistently estimable) or smaller than $\mathsf{O}(1/T)$ (as they vanish), yet $1/\sqrt{T}$ and $1/T$ both converge to zero as $T \to \infty$, so 'most' parameter values are unproblematic.

When the LDGP is not nested in the GUM, direct estimation will deliver inconsistent estimates, and while a selected approximation will also be an incorrect choice, it will be undominated, and in a progressive research strategy, especially when there are intermittent structural breaks in both relevant and irrelevant variables, will soon be replaced. Conversely, if the LDGP would always be retained by *Autometrics* when commencing from it, then a close approximation will generally be selected when starting from a GUM which nests that LDGP. Costs of inference dominate costs of search for most values of the non-centrality parameter and numbers of candidate variables. Search costs rise with the extent of initial over-specification, whereas inference costs rise with under-specification, even in constant-parameter processes. Consequently, prior theoretical analyses that can ascertain the main relevant variables and likely lag-reaction latencies remain invaluable, and can be embedded in the search process, allowing more stringent selection of other potential effects, as in Hendry and Mizon (2009).

Overall, we conclude that model selection based on *Autometrics* using relatively tight significance levels and bias correction is a successful approach to selecting dynamic equations even when commencing from very long lags to avoid omitting relevant variables or dynamics.

References

Campos, J., Hendry, D. F., and Krolzig, H.-M. (2003). Consistent model selection by an automatic Gets approach. *Oxford Bulletin of Economics and Statistics*, **65**, 803–819.

Castle, J. L., Doornik, J. A., and Hendry, D. F. (2009). Model selection when there are multiple breaks. Working paper, Economics Department, University of Oxford.

Castle, J. L., and Hendry, D. F. (2009). Automatic selection of non-linear models. Mimeo, Economics Department, Oxford University.

Castle, J. L., Qin, X., and Reed, W. R. (2009). How to pick the best regression equation: A Monte Carlo comparison of many model selection algorithms. Working paper, Economics Department, University of Canterbury, Christchurch, New Zealand.

Castle, J. L., and Shephard, N. (eds.)(2009). *The Methodology and Practice of Econometrics*. Oxford: Oxford University Press.

Clements, M. P., and Hendry, D. F. (1998). *Forecasting Economic Time Series*. Cambridge: Cambridge University Press.

Clements, M. P., and Hendry, D. F. (1999). *Forecasting Non-stationary Economic Time Series*. Cambridge, Mass.: MIT Press.

Demiralp, S., and Hoover, K. D. (2003). Searching for the causal structure of a vector autoregression. *Oxford Bulletin of Economics and Statistics*, **65**, 745–767.

Doornik, J. A. (2007a). Econometric model selection with more variables than observations. Working paper, Economics Department, University of Oxford.

Doornik, J. A. (2007b). *Object-Oriented Matrix Programming using Ox* 6th edn. London: Timberlake Consultants Press.

Doornik, J. A. (2008). Encompassing and automatic model selection. *Oxford Bulletin of Economics and Statistics*, **70**, 915–925.

Doornik, J. A. (2009). Autometrics. In Castle, and Shephard (2009), pp. 88–121.

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.

Hendry, D. F. (2009). The methodology of empirical econometric modeling: Applied econometrics

through the looking-glass. In Mills, T. C., and Patterson, K. D. (eds.), *Palgrave Handbook of Econometrics*, pp. 3–67. Basingstoke: Palgrave MacMillan.

Hendry, D. F., and Doornik, J. A. (2009). *Empirical Econometric Modelling using PcGive: Volume I*. London: Timberlake Consultants Press.

Hendry, D. F., Johansen, S., and Santos, C. (2008). Automatic selection of indicators in a fully saturated regression. *Computational Statistics*, **33**, 317–335. Erratum, 337–339.

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. (2005). The properties of automatic Gets modelling. *Economic Journal*, **115**, C32–C61.

Hendry, D. F., and Mizon, G. E. (2009). Econometric modelling of changing time series. 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.

Hoover, K. D., and Perez, S. J. (2004). Truth and robustness in cross-country growth regressions. *Oxford Bulletin of Economics and Statistics*, **66**, 765–798.

Hylleberg, S. (1986). *Seasonality in Regression*. Orlando, Florida: Academic Press.

Hylleberg, S. (ed.)(1992). *Modelling Seasonality*. Oxford: Oxford University Press.

Johansen, S., and Nielsen, B. (2009). An analysis of the indicator saturation estimator as a robust regression estimator. In Castle, and Shephard (2009), pp. 1–36.

Johnson, N. L., and Kotz, S. (1970). *Continuous Univariate Distributions*. New York: John Wiley. Volume 1.

Krolzig, H.-M. (2003). General-to-specific model selection procedures for structural vector autoregressions. *Oxford Bulletin of Economics and Statistics*, **65**, 769–802.

Kurcewicz, M., and Mycielski, J. (2003). A specification search algorithm for cointegrated systems. Discussion paper, Statistics Department, Warsaw University.

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

Leeb, H., and Pötscher, B. M. (2003). The finite-sample distribution of post-model-selection estimators, and uniform versus non-uniform approximations. *Econometric Theory*, **19**, 100–142.

Leeb, H., and Pötscher, B. M. (2005). Model selection and inference: Facts and fiction. *Econometric Theory*, **21**, 21–59.

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

Perez-Amaral, T., Gallo, G. M., and White, H. (2003). A flexible tool for model building: the relevant transformation of the inputs network approach (RETINA). *Oxford Bulletin of Economics and Statistics*, **65**, 821–838.

Phillips, P. C. B. (1994). Bayes models and forecasts of Australian macroeconomic time series. In Hargreaves, C. (ed.), *Non-stationary Time-series Analysis 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. (2003). Laws and limits of econometrics. *Economic Journal*, **113**, C26–C52.

Pötscher, B. M. (1991). Effects of model selection on inference. *Econometric Theory*, **7**, 163–185.

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

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society*, **B, 58**, 267–288.

White, H. (2000). A reality check for data snooping. *Econometrica*, **68**, 1097–1126.