Evaluating Automatic Model Selection

Jennifer L. Castle[†], Jurgen A. Doornik[‡] and David F. Hendry[⋆ *]

[†] James Martin Fellow, and Magdalen College, University of Oxford, UK

[‡] Nuffield College, University of Oxford, UK

[⋆] James Martin Fellow, and Department of Economics, University of Oxford, UK

Abstract

We outline a range of criteria for evaluating model selection approaches that have been used in the literature. Focusing on three key criteria, we evaluate automatically selecting the relevant variables in an econometric model from a large candidate set. General-to-specific selection is outlined for a regression model in orthogonal variables, where only one decision is required to select, irrespective of the number of regressors. Comparisons with an automated model selection algorithm, *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 selection in dynamic models by 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.

[*]The research was supported in part by grants from the Open Society Foundation and the Oxford-Martin School.

1 Introduction

Model selection has historically been a contentious issue, see e.g., Leamer (1978, 1983a), Chatfield (1995) and Hansen (2005). The literature has traditionally focused on situations where the only uncertainty concerns the set of relevant variables. Selection procedures such as information criteria, stepwise regression, shrinkage methods such as Lasso, or cross validation, all aim to select a set of relevant variables from a candidate set. In this paper, we consider a more general setting, whereby there is specification uncertainty over the choice of which variables, their lags, functional forms etc., are relevant and which irrelevant. This more general framework acknowledges that there is uncertainty about every aspect of model specification, and selection has the objective of reducing such uncertainties, at the cost of a 'local increase' in uncertainty at the margin of significance regarding what is relevant. However, model selection in this more complex framework is difficult: to successfully determine what matters and how it enters, all important determinants need to be included jointly. 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. To resolve that requires commencing from a sufficiently general model in which all potentially relevant variables, lags, functional forms, outliers, etc. are included initially. The objective of our paper is to consider why model selection might be successful in such a setting, and then focus on its application to dynamic specification.

The structure of the paper is as follows. Section 2 briefly describes how the paper fits in with the literature on model selection, motivating the analysis of automatic general-to-specific (*Gets*) selection. Section 3 considers nine common criteria for evaluating the success, or otherwise, of model selection. Three of those criteria are both operational and relevant to practical empirical modeling. These three criteria are applied to *Gets* selection, first for a simple 1-cut rule (section 4), then in a comparison of the 1-cut rule with the more complicated model selection procedure of automated *Gets* (section 5), and finally applying that approach to a dynamic data generation process (DGP) in section 6. Section 7 concludes.

A more detailed overview of Sections 4-6 will help to motivate their structure. Section 4 considers the analytically tractable setting of a constant DGP with orthogonal variables, an unknown subset of which are relevant to explaining the dependent variable, with the remainder being irrelevant. We show that only one selection decision is required, called '1-cut', irrespective of the number of regressors $N < T$, for T observations. This rebuts claims that model selection intrinsically involves repeated testing (see e.g. Leamer, 1983a). A simulation experiment for $N = 1000$ at $T = 2000$ confirms the theoretical analysis of '1-cut'. 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.

We introduce the terminology '*gauge*' and '*potency*'. Gauge denotes the retention frequency of irrelevant variables when selecting, without necessitating that they be 'significant'. Thus, gauge measures the 'distance' between correctly excluding all the irrelevant variables and retaining some irrelevant variables, with connotations of the gauge as the distance between rails on a railway track. Gauge is used because the 'size' of a test statistic has a definition which is only precise for a similar test with a known distribution, and the word is ambiguous in many settings (such as sample size). Similarly, retaining relevant variables no longer corresponds to the conventional notion of power, so we use potency to denote the average retention frequency of relevant variables, which need not be by rejecting the null. Potency implies strength, so is appropriate to delineate retaining variables that have 'signal strength', i.e. that are relevant.

Gauge and potency are more useful concepts than the metric used by Lovell (1983), which measures the probability that the DGP is selected as the model. That metric does not account

for variables with very low (but non-zero) population **t**-statistics, where such variables would not be retained when testing the DGP itself. For 1-cut, the 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 seeking to select a small number of relevant variables from a very large set of candidates.

Although there is no repeated testing, *selection* does affect 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 next step is to correct for biases induced in conditional distributions by only retaining significant coefficients. Building on Hendry and Krolzig (2005), we show that bias corrections also reduce mean-square errors (MSEs) of irrelevant variables estimates in both conditional and unconditional distributions, with a small increase in the MSEs of relevant variables.

Next, a more general procedure is required than 1-cut that does not depend on orthogonality of the regressors. Section 5 compares the 1-cut approach with using a general search algorithm implementing automatic *Gets*, namely *Autometrics* within *PcGive* (see Doornik, 2007b, 2009, Hendry and Doornik, 2009). The 1-cut strategy is 'optimal' when the regressors are orthogonal in-sample, so a comparison provides a check of the closeness of the more general algorithm's results. We also assess the impact of mis-specification testing on gauges, and note the role of tests for encompassing the initial general unrestricted model (GUM).

Since dynamic dependence induces correlations between adjacent lags, the resulting non-orthogonality requires a general algorithm. Section 6 extends the experimental design to stationary dynamic DGPs, examining models that are under-specified relative to the DGP, match the DGP, and are over-specified. 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). A lag-length pre-search at loose significance levels from the longest lag 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, although measurements of performance must account for the difficulty in precisely dating lag reactions. We now review previous approaches.

2 Model selection literature

There is a vast literature on model selection, including procedures based on penalized model fit: (e.g.) the C_p criterion of Mallows (1973), the prediction criterion of Amemiya (1980), and various information criteria, such as Akaike (1973), Schwarz (1978), Hannan and Quinn (1979), and Chow (1981). However, these procedures do not ensure congruency, and so a mis-specified model could be selected (see Bontemps and Mizon, 2003). Shrinkage techniques have been proposed as a solution to the 'pre-test problem': see Stein (1956), James and Stein (1961), Yancey and Judge (1976) and Judge and Bock (1978), and associated algorithms such as Lasso (Tibshirani, 1996). One criticism of shrinkage is that it is not progressive, in the sense of knowledge accumulating about the process being modelled, because the decision rule need not eliminate variables. Bayesian model averaging (see Hoeting, Madigan, Raftery and Volinsky, 1999, for an overview) is often used to account for model uncertainty, as is the extreme bounds literature of Leamer (1978, 1983b, 1985). This approach has been heavily criticised by, *inter alia*, McAleer, Pagan and Volker (1985), Breusch (1990) and Hendry and Mizon (1990). Stepwise regression is popular, but is path dependent, is susceptible to negative dependence, and does not have a high success rate of finding the DGP. Berk (1978) demonstrates that applying both forward and backward selection is no guarantee to finding the correct model. Alternative selection procedures exist, such as 'optimal regression' in which all subsets of regressors are

included (see Coen, Gomme and Kendall, 1969, and the response by Box and Newbold, 1971), but that approach is anyway intractable with a large set of potential regressors.

Given these criticisms, our paper focuses on general-to-specific model selection. The procedure commences with a large set of potential regressors and simplifies the model to capture the salient characteristics of the data: see *inter alia*, Anderson (1962), and Pagan (1987), Phillips (1988) and Campos, Ericsson and Hendry (2005) for reviews. We challenge the traditional view that model selection is costly by demonstrating the low costs of search relative to the costs of conducting inference on a pre-specified model. In contrast, the costs of search for earlier procedures can be very high. For example, selection using information criteria requires selecting from 2^N possible models, which will often be infeasible for large N, whereas *Gets* selects from the N candidate variables. Furthermore, stepwise selection could 'miss' relevant variables with negative dependencies depending on the order of inclusion, whereas *Gets* insures against path dependence by undertaking a tree search. We next consider how to evaluate model selection procedures, before focusing attention on the performance of *Gets* selection, noting that the *Gets* approach includes aspects of many other model selection methods (see §5).

3 Evaluating model selection

In this section, we consider how alternative methods of model selection might be evaluated, including the three criteria that we subsequently use. 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. Since many features of models are not derivable from subject-matter theory, empirical evidence is essential to determine what are the relevant variables, their lag reactions, parameter shifts, non-linear functions and so on. Embedding a 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. That 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 be estimable, so automatic methods for model selection are essential.

Nevertheless, the best model selection approaches cannot be expected to select the LDGP on every occasion, even when 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 in that setting.

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 relative to 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.

Criterion (A) is a traditional criterion, often based on penalized fit, but Lovell (1983) showed that it did not lead to useful selections. The second, (B), is overly demanding, as it may be nearly impossible to find the LDGP even when commencing from it, e.g., the population non-centralities of some relevant variables' coefficients may be non-zero but very small. The third criterion, (C), 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, 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), 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. Criteria (A)–(F) are all widely used but, as noted above, there are situations in which selecting by such criteria will not work well. Conversely, (G)–(I) apply to any situation and any model selection procedure, so we use (G), (H) and (I) as the basis for evaluation, noting that they could in principle be achieved together.

4 Why *Gets* model selection can succeed

In this section, we consider the simplest case of a constant-parameter linear regression model with perfectly orthogonal variables in-sample, of which a subset comprise the LDGP. This simple case is of interest because it demonstrates that model selection does not require searching through 2^N possible models. Instead, one decision is required, hence the '1-cut' rule. Then *Gets* selection can be seen as a natural generalization to the situation where the regressors are not perfectly orthogonal. We show that the 1-cut rule satifies citeria (G), (H) and (I), providing the groundwork for the more general correlated case in Section 5.

4.1 The orthogonal regressor model

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

perfectly orthogonal in sample:

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

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 $T >> N$. In (1), $n \leq N$ of the regressors have non-zero β_k, but it is not known which, nor how many.

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).

4.2 Simulation 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 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 mean square errors (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 (on retaining) MSEs respectively are calculated as:

$$\mathsf{UMSE}_k = \frac{1}{M}\sum_{i=1}^{M} \left(\beta^*_{k,i} - \beta_k\right)^2, \ \forall k \tag{5}$$

$$\mathsf{CMSE}_k = \frac{\sum_{i=1}^{M}\left[\left(\beta^*_{k,i} - \beta_k\right)^2 \cdot 1(\beta^*_{k,i} \neq 0)\right]}{\sum_{i=1}^{M} 1(\beta^*_{k,i} \neq 0)}, \quad (\beta^2_k \text{ when } \sum_{i=1}^{M} 1(\beta^*_{k,i} \neq 0) = 0) \tag{6}$$

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

GMSE_k refers to the MSE for variable k in the GUM (which is only estimable with fewer variables than observations) and LMSE_k refers to the MSE for variable k in the LDGP. After selecting from the GUM, the unconditional MSE for variable k in the resulting model is given by USMSE_k and the corresponding conditional MSE is given by CSMSE_k. If selection is

	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
LDGP	$\overline{\beta}_{k,i}$	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

undertaken commencing from the LDGP, UIMSE_k refers to the unconditional MSE for the k^{th} variable in the selected model (with I referring to 'inference costs') and correspondingly CIMSE_k is the conditional MSE. 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. §6.3 relates the costs of search and inference to Table 1.

4.3 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: the 'pre-test' problem. Conditional estimates of relevant coefficients are biased away from zero as they are only retained when $\mathsf{t}^2 \geq c_\alpha^2$, and some relevant variables will by chance have $\mathsf{t}^2 < c_\alpha^2$ in any given sample, so not be selected. Also, on average $\alpha(N-n)$ irrelevant variables will have $\mathsf{t}^2 \geq c_\alpha^2$ (spurious significance). However, bias correction after selection is easily implemented following Hendry and Krolzig (2005).

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

$$\psi^* = \mathsf{E}\left[\mathsf{t}_{\widetilde{\beta}} \mid \left|\mathsf{t}_{\widetilde{\beta}}\right| > c_\alpha; \psi\right] = \psi + \frac{\phi(c_\alpha - \psi) - \phi(-c_\alpha - \psi)}{1 - \Phi(c_\alpha - \psi) + \Phi(-c_\alpha - \psi)} = \psi + r(\psi, c_\alpha) \tag{7}$$

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

$$\mathsf{E}\left[\widetilde{\beta} \mid \widetilde{\beta} \geq \sigma_{\widetilde{\beta}} c_\alpha\right] = \beta + \sigma_{\widetilde{\beta}} r(\psi, c_\alpha) = \beta\left(1 + \psi^{-1} r(\psi, c_\alpha)\right) \tag{8}$$

so an unbiased estimator after selection is:

$$\widetilde{\widetilde{\beta}} = \widetilde{\beta}\left(\frac{\psi}{\psi + r(\psi, c_\alpha)}\right) = \widetilde{\beta}\left(\frac{\psi}{\psi^*}\right). \tag{9}$$

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 (7) written as:

$$\psi = \psi^* - r(\psi, c_\alpha) \tag{10}$$

First replace $r(\psi, c_\alpha)$ in (10) by $r(\mathsf{t}_{\widetilde{\beta}}, c_\alpha)$, and ψ^* by $\mathsf{t}_{\widetilde{\beta}}$:

$$\widetilde{\mathsf{t}}_{\widetilde{\beta}} = \mathsf{t}_{\widetilde{\beta}} - r\left(\mathsf{t}_{\widetilde{\beta}}, c_\alpha\right), \text{ then } \widetilde{\widetilde{\mathsf{t}}}_{\widetilde{\beta}} = \mathsf{t}_{\widetilde{\beta}} - r\left(\widetilde{\mathsf{t}}_{\widetilde{\beta}}, c_\alpha\right) \tag{11}$$

leading to the bias-corrected parameter estimate:

$$\widetilde{\widetilde{\beta}} = \widetilde{\beta}\left(\widetilde{\widetilde{\mathsf{t}}}_{\widetilde{\beta}}/\mathsf{t}_{\widetilde{\beta}}\right). \tag{12}$$

Hendry and Krolzig (2005) show that most of the selection bias is corrected for relevant retained variables by (12), 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. The power loss from tighter significance levels is usually not substantial relative to, say, a t-distribution with few degrees of freedom. However, Castle, Doornik and Hendry (2010) show that impulse-indicator saturation (see Hendry, Johansen and Santos, 2008, and Johansen and Nielsen, 2009) is a successful antidote for fat-tailed error processes.

4.4 Monte Carlo simulation of 1-cut for $N = 1000$

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

$$\begin{aligned} y_t &= \beta_1 x_{1,t} + \cdots + \beta_{10} x_{10,t} + \epsilon_t && (13)\\ \mathbf{x}_t &\sim \mathsf{IN}_{1000}\left[\mathbf{0}, \mathbf{I}\right] && (14)\\ \epsilon_t &\sim \mathsf{IN}\left[0, 1\right] && (15) \end{aligned}$$

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 t-test non-centralities, ψ_k, are reported in Table 2, together with the theoretical powers of t-tests on the individual coefficients.

	$k=1$	$k=2$	$k=3$	$k=4$	$k=5$	$k=6$	$k=7$	$k=8$	$k=9$	$k=10$
β_k	0.063	0.079	0.095	0.111	0.126	0.142	0.158	0.174	0.190	0.206
ψ_k	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.

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

$$y_t = \beta_0 + \beta_1 x_{1,t} + \cdots + \beta_{1000} x_{1000,t} + u_t, \quad t = 1, \ldots, 2000 \tag{16}$$

Only the first $n = 10$ variables are relevant, so 991 variables are irrelevant in (16). Selection is undertaken by estimating (16), 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,

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

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

even when selecting 10 relevant regressors from 1000 candidate variables in one decision. Figure 1 confirms that 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 estimated coefficients in (16) are 0.001 as shown.

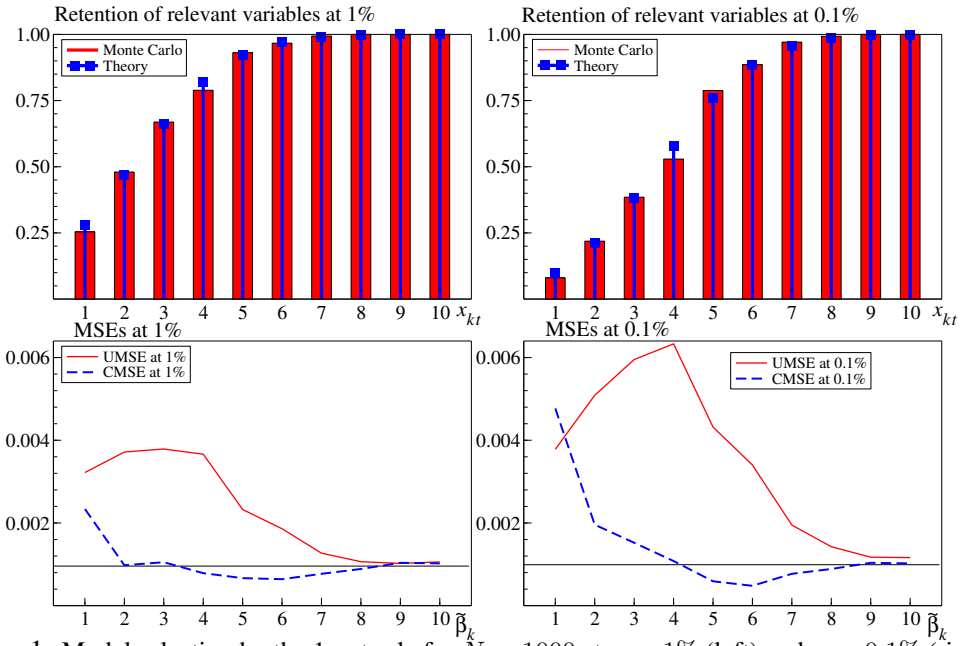

Figure 1: Model selection by the 1-cut rule for $N = 1000$ at $\alpha = 1\%$ (left) and $\alpha = 0.1\%$ (right): retention rates $\widetilde{p}_k$ of relevant variables $x_1, \ldots, x_{10}$ (top graphs), USMSE_k and CSMSE_k (bottom graphs).

Figure 2 records the trade-off frontier of gauge against potency as the non-centrality increases. This can be compared to the theoretical frontier based on a single t-test for an individual coefficient, recording size against power. The difference between the two frontiers is very small, due to sampling variation, demonstrating that the 1-cut algorithm matches the theory. See Hoover and Perez (1999, figure 1) for a similar analysis of the potency/gauge frontier.

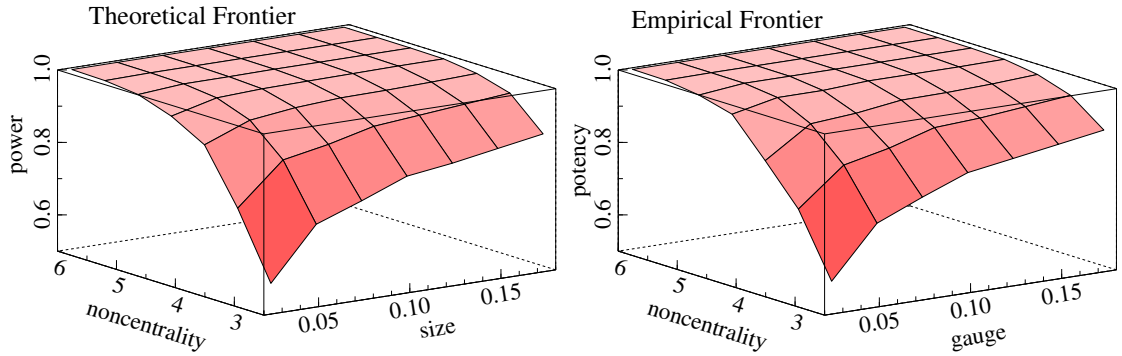

Figure 2: Trade-off frontier: left-hand panel records the theoretical frontier of size against power; right-hand panel records the empirical frontier of gauge against potency.

4.5 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 their deletion would make a diagnostic test significant, or because of encompassing since a variable can be individually insignificant, but not jointly with all variables deleted so far. The bias correction formula in (12) is only applied 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 eliminated at $\alpha = 0.001$, their USMSEs are tiny. Thus, in complete contrast to the earlier literature reviewed in section 2, even with 991 irrelevant variables, their total impact on selected models after bias correction is essentially negligible when suitable significance levels are used. Figure 3 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.

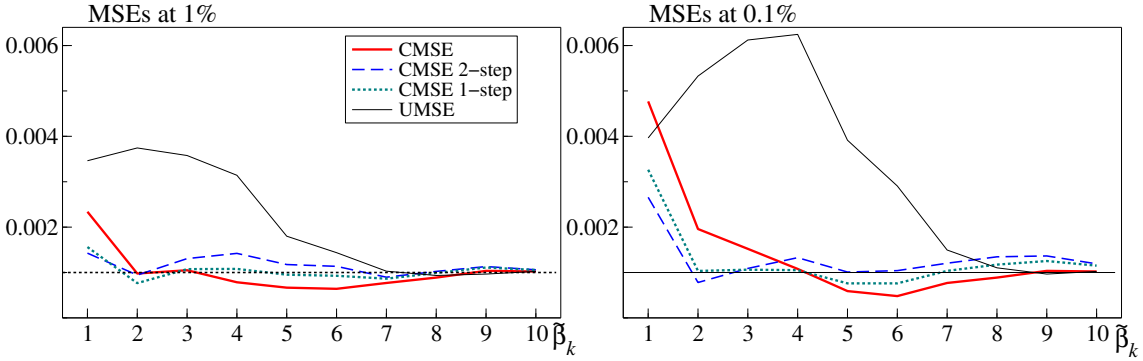

Figure 3: Impact of bias correction on CSMSE_k for relevant variables at $\alpha = 1\%$ (left) and $\alpha = 0.1\%$ (right).

5 Comparisons of 1-cut selection and automated *Gets*

Having demonstrated that the 1-cut rule satisfies properties (G), (H) and (I), we now compare it with automated *Gets*. 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. The multi-path procedures in Hoover and Perez (1999) and Hendry and Krolzig (2001) do 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). *Autometrics* improves further by a tree-search to detect and eliminate statistically-insignificant variables, and handling $N > T$. At any

stage, a variable is removed only 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, a choice is made 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).

Gets selection encompasses many aspects of the alternative approaches discussed in §2. For example, information criteria are used to select a final model if there are multiple terminal models. Alternatively, the terminal models could be averaged, or the GUM could be retained, or included in a set of models being averaged. The GUM should be specified using all available subject-matter theory, and such theory can be imposed by 'forcing' variables to be retained. When there are more variables than observations, expanding and contracting searches are used, as in Doornik (2009), so specific-to-general search also plays a role. Encompassing tests ensure a parsimonious congruent model of the GUM is selected, see Doornik (2008).

We now consider a much smaller N 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). We retain the orthogonal design to ensure that 1-cut is a valid procedure. The analysis shows that there is little or no cost to undertaking a multi-path search compared to 1-cut, validating the *Gets* procedure.

5.1 A small orthogonal-regressor model

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

$$
\begin{aligned}
y_t &= \beta_0 + \beta_1 x_{1,t} + \cdots + \beta_{10} x_{10,t} + \epsilon_t, && (17)\\
\mathbf{x}_t &\sim \mathsf{IN}_{10}\left[\mathbf{0}, \mathbf{I}_{10}\right], && (18)\\
\epsilon_t &\sim \mathsf{IN}\left[0, \sigma_j^2\right],\ t = 1, \ldots, T, \quad j = 2, \ldots, 6, && (19)
\end{aligned}
$$

where $\mathbf{x}_t' = (x_{1,t}, \cdots, x_{10,t})$. The $\mathbf{x}_t$ are fixed across replications. Equations (17)–(19) 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.

The error variance is given by σ_j^2 where j indexes 5 different error variances calculated by $\sigma_j^2 = T/j^2$, for $j = 2, \ldots, 6$, such that all relevant variables in each experiment have the same non-centrality given by $\psi_{k,j} = 2, \ldots, 6$ for $k = 1, \ldots, n$. Hence, the 10 different DGPs have relevant variables with 5 different non-centralities, resulting in 50 experiments. The experimental design aims to span a broad range of situations, from many relevant regressors to few relevant regressors, and from highly significant to marginally significant regressors, to ensure the simulation results are relatively general within the linear, orthogonal-regressor context. Table 5 reports the theoretical powers of t-tests for the non-centralities of relevant variables considered.

The GUM is the same for all 10 DGPs:

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

	$j=2$	$j=3$	$j=4$	$j=5$	$j=6$
$\alpha=0.05$	50.3	84.3	97.8	99.9	100
$\alpha=0.01$	26.0	63.9	91.3	99.1	100

Table 5: Theoretical power for a single t-test (%) for experiments (17)–(19).

5.1.1 Simulation results for $N = 10$

We now investigate how the general search algorithm performs relative to 1-cut selection in terms of (G)–(I) in Section 3. Their comparative gauges for $\psi_k = 2$ and $\psi_k = 6$ are shown in Figure 4, where *Autometrics* selects both with and without diagnostic testing. In default mode (with diagnostic testing), *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 §5.2). Doornik (2008) shows that encompassing checks against the GUM help stabilize performance.

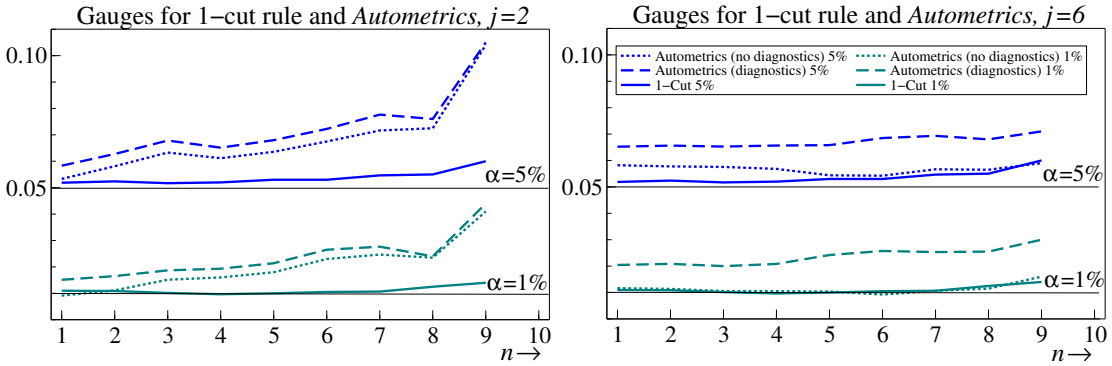

Figure 4: Gauges for 1-cut rule (solid lines), *Autometrics* with diagnostic testing (dashed lines) and *Autometrics* without diagnostic testing (dotted lines) for $\alpha = 0.01, 0.05$. The left panel corresponds to $j = 2$ so $\psi_k = 2$ for $k = 1, \ldots, n$, and the right panel corresponds to $j = 6$, so $\psi_k = 6$ for $k = 1, \ldots, n$. The horizontal axis represents the $n = 1, \ldots, 10$ DGPs, each with n relevant variables (and $10 - n$ irrelevant).

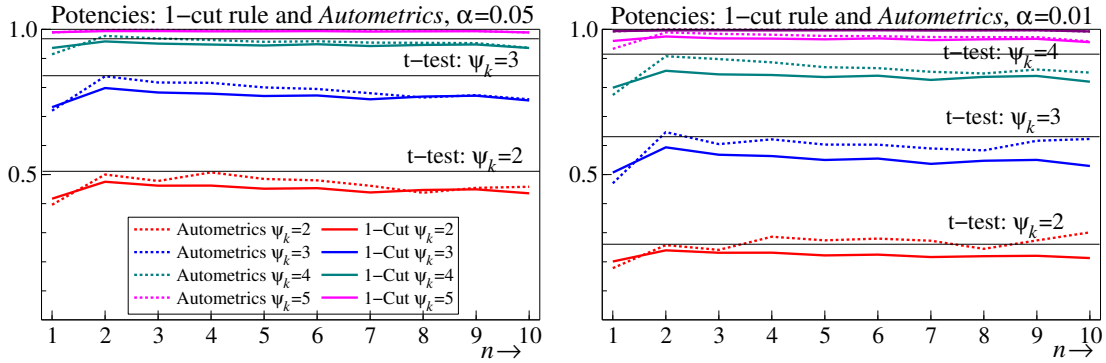

Figure 5: Potencies for 1-cut rule (solid lines) and *Autometrics* without diagnostic testing (dotted lines) for $\alpha = 0.05$ (left panel) and 0.01 (right panel). The horizontal axis represents the $n = 1, \ldots, 10$ DGPs, each with n relevant variables (and $10 - n$ irrelevant). Solid thin lines record the power for a single t-test at $\psi_k = 2, 3, 4$.

Figure 5 compares the potencies of both algorithms, without diagnostic testing, and when the intercept is always included. Potencies can be compared to the power of a single t-test, also recorded in Figure 5 (powers for high non-centralities are excluded as they are close to unity). Both methods have potencies close to the optimal single t-test with no selection. The 1-cut rule has a consistently lower potency, but potencies are not gauge-corrected, and it also has a

slightly lower gauge. Given this trade-off, there is little difference between 1-cut and searching many paths.

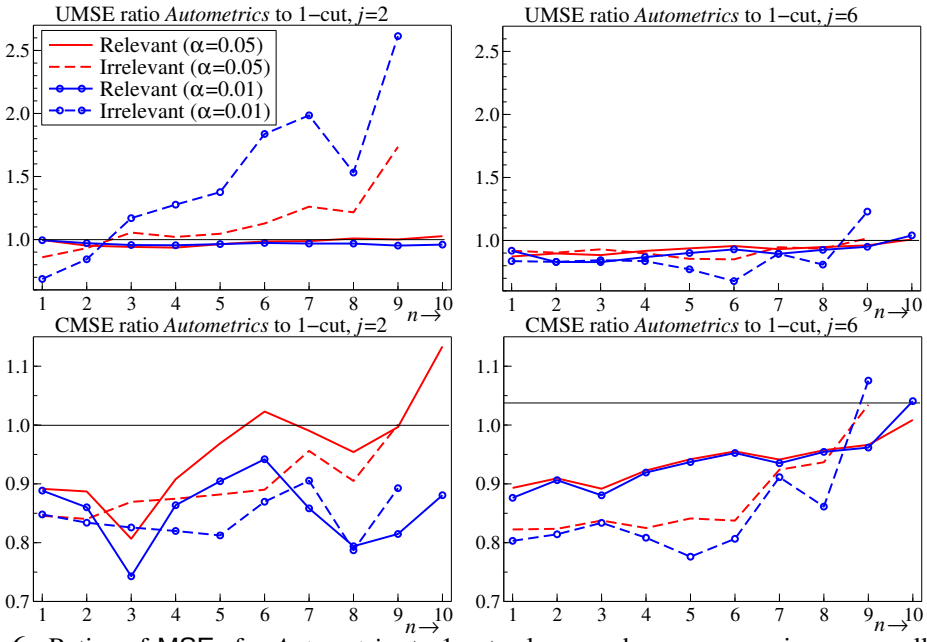

Figure 6: Ratios of MSEs for *Autometrics* to 1-cut rule as n changes, averaging across all relevant (solid lines) and irrelevant (dashed lines) variables. Left-hand panels correspond to $j = 2$ ($\psi_k = 2$) and right-hand panels correspond to $j = 6$ ($\psi_k = 6$) for $k = 1, \ldots, n$.

Figure 6 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, the former 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, all ratios are close to unity for the relevant variables, but the 1-cut rule performs better for irrelevant variables when non-centralities are low but not when they are high. The benefits to search are largest when there are few relevant variables that are highly significant, but conditionally, *Autometrics* outperforms the 1-cut rule in almost all cases–most lines are below unity. Thus, there is little loss from using the path-search algorithm even when 1-cut is applicable. In non-orthogonal problems, 1-cut would be inadvisable as the initial ranking given by (2) depends on correlations between variables as well as their relevance.

5.2 Impact of diagnostic tests

Figure 4 also compared the gauges for *Autometrics* with diagnostic tracking switched on versus off, both with bias correction. The gauge is slightly over the nominal significance level when diagnostic tests are checked to ensure a congruent reduction. With diagnostic testing 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–a key reason for measuring gauge not 'size'.

Figure 7 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. 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.

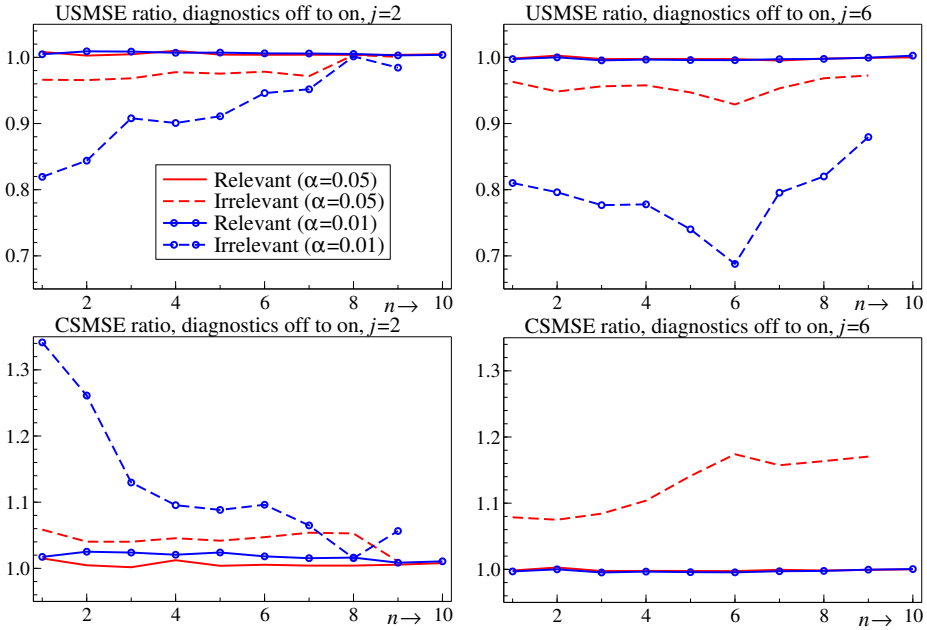

Figure 7: Ratios of MSEs with diagnostic tests off to on for unconditional and conditional distributions as n changes, averaging across all relevant (solid lines) and irrelevant (dashed lines) variables. Left-hand panels correspond to $j = 2$ ($\psi_k = 2$) and right-hand panels correspond to $j = 6$ ($\psi_k = 6$) for $k = 1, \ldots, n$.

6 Model selection in ADL models

In this section, we analyze a setting in which the regressors are correlated and the models are dynamic; the setting that confronts most time-series researchers, so 1-cut is an invalid procedure. We use a relatively general experimental design to cover many possible settings, 56 experiments in total, for a linear DGP, and assess performance based on criteria (G)–(I). Non-linear model selection is discussed in Castle and Hendry (2010a, 2010b).

6.1 The dynamic model

The experimental design has nine DGP specifications given by, for $r = 3, \ldots, 8$:

$$\begin{aligned} \text{DGP0}: \quad y_t &= \epsilon_t \\ \text{DGP1}: \quad y_t &= 0.75 y_{t-1} + \epsilon_t \\ \text{DGP2}: \quad y_t &= 1.5 y_{t-1} - 0.8 y_{t-2} + \epsilon_t \end{aligned}$$

$$\text{DGPr}: \quad y_t = 1.5 y_{t-1} - 0.8 y_{t-2} + \sum_{j=1}^{r-2} \left(\beta_j x_{j,t} - \beta_j x_{j,t-1} \right) + \epsilon_t \tag{20}$$

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

$$\mathbf{x}_t = \rho \mathbf{x}_{t-1} + \mathbf{v}_t \text{ where } \mathbf{v}_t \sim \mathsf{IN}_6[\mathbf{0}, \mathbf{\Omega}] \tag{21}$$

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 $s = 0, 1, 2, 5, 10, 15, 20$:

$$y_t = \mu + \sum_{k=1}^{s} \alpha_k y_{t-k} + \sum_{j=1}^{6} \sum_{k=0}^{s} \gamma_{j,k} x_{j,t-k} + e_t. \tag{22}$$

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

6.2 Potency and gauge

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

DGP	3	4	5	6	7	8
ψ_k	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 8 records the potencies for each DGP and GUM specification defined by the lag length, s, commencing at $s = 2$ when there is no under-specification, for selection using lag pre-search. There is a decline in potency as the non-centrality falls (i.e., the DGP size 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 (LDVs) are high, so even including 20 lags of y has little effect on potency.

The differences between significance levels are fairly small as the lagged dependent variables have potencies close to unity. 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 9 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, the gauge is close to the nominal significance level and is somewhat tighter for moderate lag lengths. For the DGPs with just lagged dependent variables (DGP1 and DGP2), the gauges are also close to the nominal significance level, and additional lags do

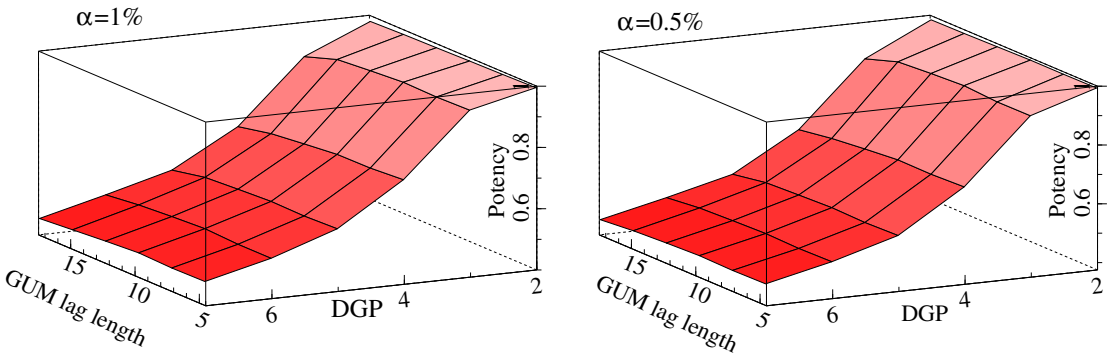

Figure 8: Potency with lag pre-search recorded against DGP and GUM specification

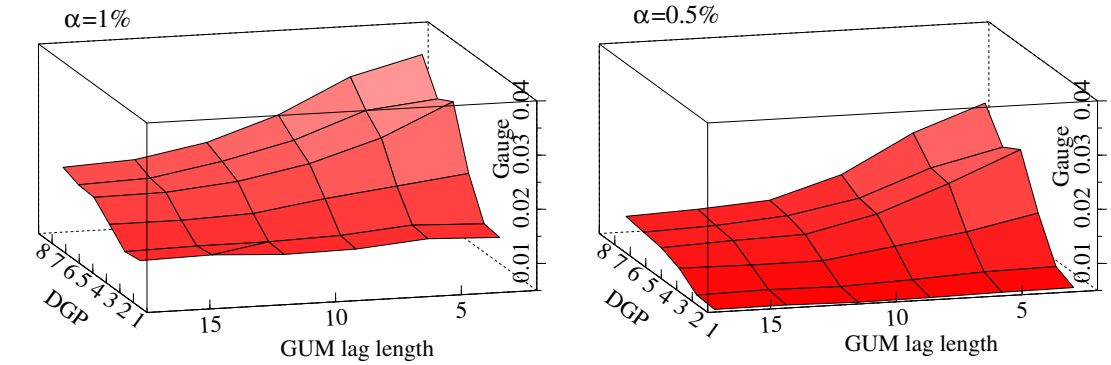

Figure 9: Gauge with lag pre-search recorded against DGP and GUM specification

not increase them. The gauges increase as more exogenous regressors become relevant, but 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 expanding and contracting searches are needed to reduce the GUM to an estimable intermediate model, see Doornik (2007a). Despite commencing with $N > T$, the gauge is controlled close to the nominal significance level. Overall, divergences from a flat plane are not substantial, so (H) seems to be satisfied even with dynamics.

6.3 Costs of search and inference

We next consider costs of inference and costs of search to assess the selection procedure, addressing criterion (G). Search and inference costs are more useful concepts than (A) or (B), because they account for the costs of conducting inference on the LDGP itself. If the signal-to-noise ratio of the LDGP is low, even commencing from it will result in some relevant variables being excluded. Such inference costs are an unavoidable consequence of the LDGP specification. We measure them by the RMSEs of LDGP parameter estimates after conducting inference on that LDGP, summing the unconditional RMSEs over all variables, namely (see Table 1 for definitions):

$$\sum_{k=1}^{n} \mathsf{UIRMSE}_k. \tag{23}$$

When the GUM is the LDGP, as only significant variables are retained, (23) could be larger or smaller than the RMSE from directly estimating the LDGP, calculated by summing the unconditional RMSEs over all variables in the estimated LDGP, depending on the choice of

critical value, c_α, and the non-centralities, ψ_k, of the LDGP parameters:

$$\sum_{k=1}^{n} \mathsf{LRMSE}_k. \tag{24}$$

Now consider starting from a more general model with both relevant and irrelevant variables. The additional costs of search are calculated as the increase in unconditional RMSEs for relevant variables in the selected model when starting from the GUM as against the LDGP, plus the unconditional RMSEs computed for all $N - n$ irrelevant variables, both bias corrected:

$$\sum_{k=1}^{n} (\mathsf{USRMSE}_k - \mathsf{UIRMSE}_k) + \sum_{k=n+1}^{N} \mathsf{USRMSE}_k \tag{25}$$

If the LDGP specification is known and just estimated, then $N = n$ and (25) is zero. Otherwise, depending on the non-centralities, ψ_k, there is usually a trade-off between the two components: as c_α increases, the second term falls, but the first term 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 6.5 considers under-specified models, where (25) can be smaller than (23), 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 (25). 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:

$$\mathsf{USRMSE}_{\mathsf{LDV}} = \sqrt{\frac{1}{M}\sum_{i=1}^{M}\left(\sum_{k=1}^{s}\widetilde{\beta}_{y,k,i} - \sum_{k=1}^{s}\beta_{y,k,i}\right)^2} \tag{26}$$

where $\widetilde{\beta}_{y,k,i}$ denotes the OLS estimate of the k^{th} 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 (26) to the costs of inference, which also sum the retained lagged dependent variables' coefficients when commencing from the LDGP.

Figure 10 records RMSEs for the LDGP given by (24), the costs of inference given by (23) and the costs of search given by (25) 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 large DGPs) despite a shift from $N \ll T$ to $N > T$ between $s = 10$ and $s = 15$. 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 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 up to $s = 5$). 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

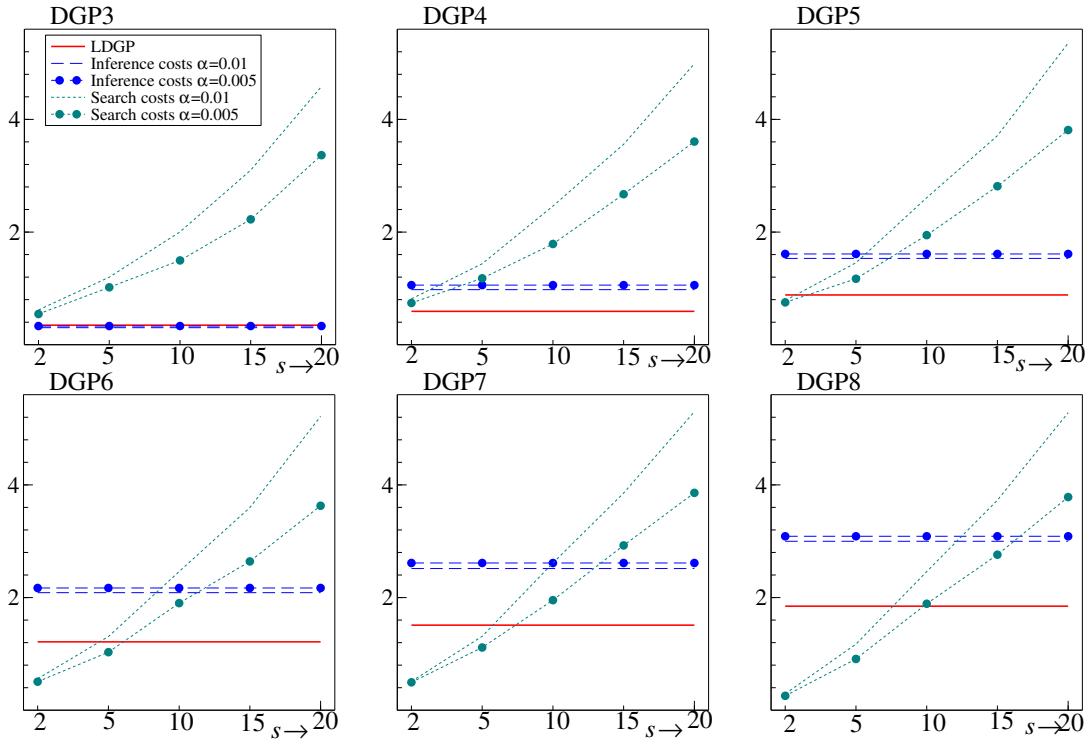

Figure 10: Costs of search and inference for exogenous regressors: LRMSE for the LDGP (solid line), inference on the LDGP (dashed lines) and bias-corrected selection from the GUM with lag pre-selection (dotted lines).

additional 98 irrelevant variables are searched over. The costs of inference over the LDGP with no selection are substantial for the larger DGPs, due to the lower non-centralities.

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 be out only 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, may be incorrect, but not their overall effect.

6.4 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 11 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 variables, 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 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 $N > T$.

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, but is infeasible when $N > T$. Computing ratios of USMSEs and CSMSEs with and without lag pre-search results in values close to unity,

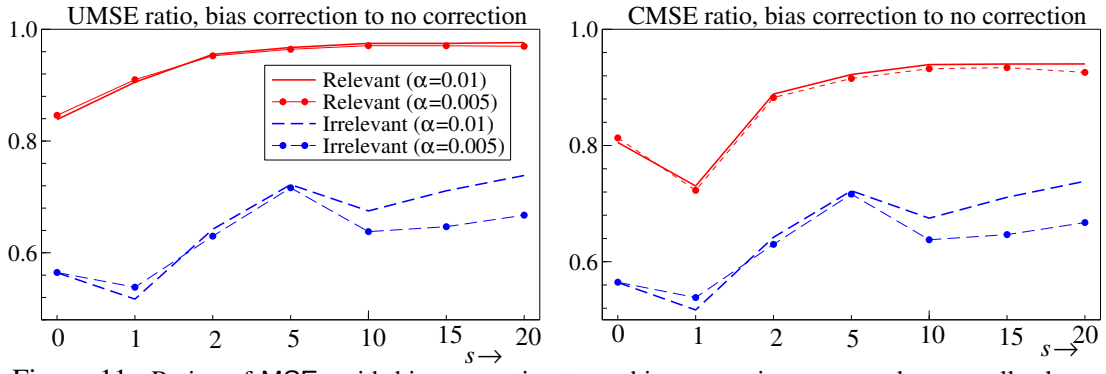

Figure 11: Ratios of MSEs with bias correction to no bias correction, averaged across all relevant (solid lines) and irrelevant (dashed lines) for all DGPs, plotted against s.

although search time is vastly improved with lag pre-selection. There is a small benefit to lag pre-search when the GUM specification includes five lags for the irrelevant variables (detailed results available on request).

6.5 Under-specification

The GUM is under-specified for all DGPs when $s = 0$, and for DGP2–DGP8 when $s = 1$. 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 selection by how often that under-specified representation is found sheds little light on how useful that 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 evaluation remains the DGP parameters, not the induced parameters of the LDGP.

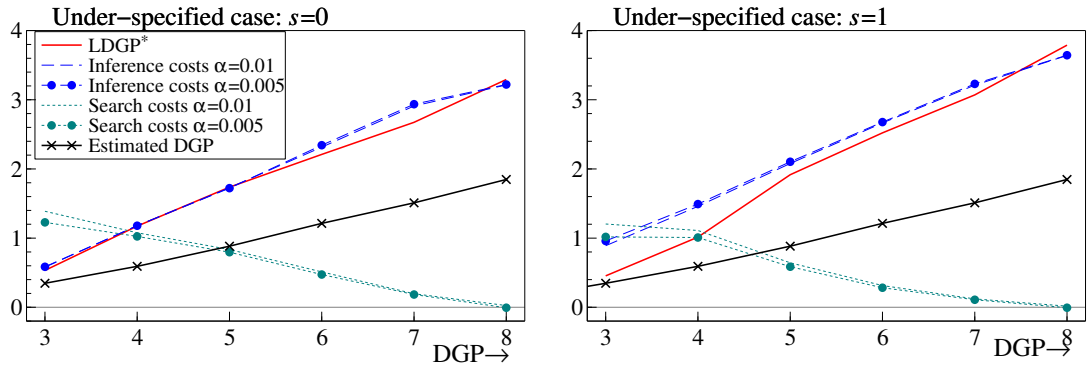

Figure 12: Costs of search and inference for under-specified cases. URMSEs for the LDGP* (solid line), inference on the LDGP* (dashed lines) and bias-corrected selection from the GUM with lag pre-selection (dotted lines).

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 12 records the DGP and LDGP* costs from (24), as well as the costs of inference, (23), and the costs of search, (25), 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 best specification. The GUMs for DGP1 and DGP2 are just a constant for $s = 0$, so are omitted from Figure 12.

6.6 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 (22) 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. The gauge, potency and search costs were close to those found for the experiments above, so 'gaps' in the dynamics have little impact: performance on such 'seasonally dynamic' data is as good as on non-seasonal data.

7 Conclusion

In this paper, we consider how to evaluate model selection procedures. Three criteria are highlighted as useful benchmarks to assess any method of model selection, namely, the local data generation process (LDGP) can be recovered when commencing from the initial general unrestricted model (GUM) almost as often as as when commencing from the LDGP itself; the operating characteristics of the selection procedure match their desired properties; and the method finds a well-specified, undominated model of the LDGP. Using these three objective assessment criteria, we examine automatic *Gets* selection embodied in *Autometrics*, see Doornik (2009). The analysis builds from the simple case of a linear regression model with N orthogonal variables, where only one decision is required to select when $N < T$ (a 1-cut procedure), to the more realistic setting of dynamic models with long lag structures and intercorrelated variables. As the automatic selection algorithm is complicated, we undertake a range of simulation experiments to assess its properties, including cases where $N > T$, and where the GUM is under-specified for the LDGP.

When the nominal rejection frequency of individual selection tests, α, is set at $\alpha \leq 1/N$, then on average one irrelevant variable will be spuriously retained as 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 retention frequency of irrelevant variables (gauge) close to α, somewhat increased by undertaking mis-specification testing for congruence and encompassing tests against the GUM. Bias correction for selection greatly reduces the mean square errors (MSEs) of spuriously 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 in N despite $N > T$. Thus, the procedure usually

terminates with a selected model close to what would be found commencing from the LDGP, with near unbiased estimates of the retained LDGP parameters, and almost no irrelevant variables, with those retained having small MSEs, thereby satisfying our three criteria.

Limits to 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 cannot rectify that. When relevant variables have parameters that are $\mathsf{O}(1/\sqrt{T})$, and regressors are highly intercorrelated, selection will not work well, see Leeb and Pötscher (2003, 2005). Thus, uniform convergence seems infeasible, as parameters cannot then be consistently estimated. However, 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), and $1/\sqrt{T}$ and $1/T$ both converge to zero as $T \to \infty$, so 'most' parameter values are unproblematic. The so-called 'size' of a selection procedure, $1-(1-\alpha)^{N-n}$, can be large, but is uninformative about the success of selection that correctly eliminates $(1-\alpha)(N-n)$ irrelevant variables on average, and is consistent when $\alpha \to 0$ as $T \to \infty$.

When the LDGP is not nested in the GUM, direct estimation will deliver inconsistent estimates. 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 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 (2010). Automatic model selection is just the next step up from automatic computation, extending the capabilities of empirical modellers.

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

Akaike, A. (1973). Information theory and an extension of the maximum likelihood principle. In Petrov, B. N., and Csaki, F. L. (eds.), *Second International Symposium of Information Theory*, pp. 267–281. Budapest: Akademiai Kiado.

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

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

Berk, K. N. (1978). Comparing subset regression procedures. *Technometrics*, **20**, 1–6.

Bontemps, C., and Mizon, G. E. (2003). Congruence and encompassing. In Stigum, B. P. (ed.), *Econometrics and the Philosophy of Economics*, pp. 354–378. Princeton: Princeton University Press.

Box, G. E. P., and Newbold, P. (1971). Some comments on a paper of Coen, Gomme, and Kendall. *Journal of the Royal Statistical Society, A*, **134**, 229–240.

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

Campos, J., Ericsson, N. R., and Hendry, D. F. (eds.)(2005). *Readings on General-to-Specific Modeling*. Cheltenham: Edward Elgar.

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. (2010). Model selection when there are multiple breaks. Working Paper No. 472, Economics Department, University of Oxford.

Castle, J. L., and Hendry, D. F. (2010a). Automatic selection of non-linear models. In Wang, L., Garnier, H., and Jackman, T. (eds.), *System Identification, Environmental Modelling and Control*, forthcoming. New York: Springer.

Castle, J. L., and Hendry, D. F. (2010b). A low-dimension, portmanteau test for non-linearity. *Journal of Econometrics*, **158**, 231–245.

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.

Chatfield, C. (1995). Model uncertainty, data mining and statistical inference. *Journal of the Royal Statistical Society, A*, **158**, 419–466. With discussion.

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

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

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

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.

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

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

Hansen, B. E. (2005). Challenges for econometric model selection. *Econometric Theory*, **21**, 60–68.

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. (1990). Procrustean econometrics: or stretching and squeezing data. In Granger (1990), pp. 121–136.

Hendry, D. F., and Mizon, G. E. (2010). Econometric modelling of time series with outlying observations. *Journal of Time Series Econometrics*, this issue.

Hoeting, J. A., Madigan, D., Raftery, A. E., and Volinsky, C. T. (1999). Bayesian model averaging: A tutorial. *Statistical Science*, **14**, 382–417.

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.

James, W., and Stein, C. (1961). Estimation with quadratic loss. In Neyman, J. (ed.), *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, pp. 361–379. Berkeley: University of California 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.

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

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

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

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

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

Leamer, E. E. (1985). Sensitivity analyses would help. *American Economic Review*, **75**, 308–313.

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.

Mallows, C. L. (1973). Some comments on C_p. *Technometrics*, **15**, 661–675.

McAleer, M., Pagan, A. R., and Volker, P. A. (1985). What will take the con out of econometrics?. *American Economic Review*, **95**, 293–301.

Pagan, A. R. (1987). Three econometric methodologies: a critical appraisal. *Journal of Economic Surveys*, **1**, 3–24.

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. (1988). Reflections on econometric methodology. *Economic Record*, **64**, 344–359.

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.

Stein, C. (1956). *Inadmissibility of the usual estimator for the mean of a multivariate normal distribution*. Berkeley: University of California Press.

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.

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