ISSN 1471-0498

DEPARTMENT OF ECONOMICS

DISCUSSION PAPER SERIES

MODEL SELECTION WHEN THERE ARE MULTIPLE BREAKS

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

Number 407
October 2008

Manor Road Building, Oxford OX1 3UQ

Model Selection when there are Multiple Breaks

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

Abstract

We consider selecting an econometric model when there is uncertainty over both the choice of variables and the occurrence and timing of multiple location shifts. The theory of general-to-simple (Gets) selection is outlined and its efficacy demonstrated in a new set of simulation experiments first for a constant model in orthogonal variables, where only one decision is required to select irrespective of the number of regressors (less than the sample size). That generalizes to including an impulse indicator for every observation in the set of candidate regressors (impulse saturation), as analyzed by Hendry, Johansen and Santos (2008) and Johansen and Nielsen (2009). Monte Carlo experiments show its capability of detecting up to 20 shifts in 100 observations.

JEL classifications: C51, C22.

'Any sufficiently advanced technology is indistinguishable from magic.' Arthur C. Clarke, *Profiles of The Future*, 1961

1 Preface by David Hendry

It is a pleasure and privilege to contribute to this volume in honor of Peter Phillips. Peter's publications are notable by the power and generality of the ideas, combined with clear explanations and an incredible span of the entire discipline, setting a standard that few can achieve. His major advances across so many areas are one of the reasons for the rapid progress in our discipline. Peter and I first met when Peter was a doctoral student at LSE working with Denis Sargan–Denis had also been my supervisor and was a major inspiration to both of us. Peter was one of a series of distinguished New Zealanders to come to the LSE, from his namesake, Bill Phillips (who had taught me), through Rex Bergstrom and Cliff Wymer. From the outset, it was clear that Peter had a deep commitment to econometrics, although his actual impact has far exceeded even the LSE faculty's greatest hopes. Peter has constructed a mighty edifice of invaluable results on the base bequeathed by earlier econometricians, resolving one intractable time-series problem after another. Peter has investigated estimation, distribution, inference, identification, specification, selection and forecasting, for finite and large samples, in discrete and continuous time, frequency and time domain, stationary and non-stationary processes, linear and non-linear, Bayesian and classical, analytical and computational, advancing the toolkit in too many ways to even list. Peter has also made numerous important professional contributions, including producing an entire generation of doctoral students whose combined output is truly vast; creating *Econometric Theory* and raising it to a position of pre-eminence; and helping revitalize and record the history of our discipline in an invaluable archive of interviews with its pioneers. Quietly spoken, but quick as lightning, he has enlivened many conferences, always moving understanding forward, often more than the speaker. Econometrics is in a far stronger state today from the hundreds of ideas and results that Peter has produced directly, and the thousands that have flowed indirectly: we all three wish him continuing high productivity, and cannot imagine him 'retiring'.

*Financial support from the ESRC under Research Grant RES-062-23-0061 is gratefully acknowledged.

2 Introduction

Our contribution concerns modeling situations that involve specification uncertainty over the choice of which variables, lags, functional forms etc., are relevant and which irrelevant, jointly with determining the occurrence and timing of multiple breaks affecting the model. To successfully determine what matters and how it enters, all potential 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. However, the 'Catch 22' is that there are then bound to be more variables N in total than the number of observations T, so all cannot be entered from the outset. To resolve this conundrum, our analysis proceeds in three stages.

We first provide an explanation of why general-to-simple (*Gets*) selection is efficient for a constant model in orthogonal variables, and demonstrate that only one selection decision is required irrespective of the number of regressors $N < T$. A new set of simulation experiments for $N = 1000$ confirms the theoretical analysis underlying what we call the 1-cut approach. Thus, although there are 10^{301} possible models, only 1 needs estimated and **one** decision is required, so 'repeated testing' does not occur, retention rates for irrelevant variables are close to the nominal significance level, α, for small α (e.g., $\alpha \leq 1/N$) which can be controlled, and retention rates for relevant variables are close to the theoretical power for a one-off test. Selection *per se* 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), so we correct for the biases induced in the conditional distributions, and show that on balance this can also improve the unconditional distributions as measured by average mean-square errors (MSEs).

Next, the 1-cut approach is contrasted with the outcome using a general search algorithm that does not depend on orthogonality of regressors, here *Autometrics*, an Ox Package (see Doornik, 2007, 2009) implementing automatic *Gets*, to show the closeness of their outcomes. We also assess the additional cost of mis-specification testing through its impact on false null retention rates, and the benefits of bias-correction on MSEs. To illustrate, we also use a much smaller N (namely 10), so the results can be graphed and compared.

Since large numbers of candidate regressors are unproblematic for $N << T$, we then consider the setting where $N > T$, which will automatically arise in our approach to detecting and removing multiple shifts by including an impulse indicator for every observation in the set of regressors (called impulse saturation, analyzed by Hendry *et al.*, 2008, and Johansen and Nielsen, 2009). When the T impulses are combined with the other regressors, N must exceed the sample size, so impulses are entered in large blocks, two sets of $T/2$ in the basic theory, but in smaller combinations in practice, so both simplification and expansion steps are in fact used. Johansen and Nielsen (2009) prove that under the null of no outliers or shifts, there is almost no loss of efficiency in testing for T impulses for $\alpha \leq 1/T$, even in dynamic models. While surprising at first sight, retaining an impulse when it is not needed merely removes one observation, which is all that will happen on average. Thus, efficiency is of the order of $(1 - \alpha)\,\%$. Monte Carlo experiments have confirmed the null distribution, and here we show that impulse saturation is capable of detecting up to 20 outliers in 100 observations as well as multiple shifts, including breaks close to the start and end of the sample. We compare our approach with step-wise regression–which can also handle $N > T$–and demonstrate the major gains from *Autometrics*, and consider a fat-tailed distribution.

The structure of the paper is as follows. Section 3 first discusses how to evaluate model selection approaches. Then section 4 discusses the 1-cut theory and simulation findings, followed in section 5 by the corresponding Monte Carlos for *Autometrics*. Section 6 sketches the theory of impulse saturation and section 7 reports a series of Monte Carlo experiments examining its ability to detect various forms and numbers of shifts, as well as comparing with step-wise regression. Section 8 concludes.

3 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 variables, 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 one selected the 'correct model' from some set of candidate variables that nested the LDGP, which essentially assumes the 'axiom of correct specification' for the proposed model. The 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. Phillips (2003) provides an insightful analysis of the limits of econometrics.

An automatic method offers a number of advantages:

1 Speed–with 100 candidate regressors there are too many combinations, or search paths, to explore manually;

2 Numerosity–very general models can be too large for humans to understand or manipulate;

3 Complexity–multiple breaks, non-linearities, dynamics, systems, exogeneity, integrability, interactions, etc., all need addressed simultaneously;

4 Expertise–can build in 'best practice' knowledge or an 'expert' framework;

5 Objectivity and replicability–should find the same outcome given the same starting point and selection criteria.

The first is simply the next stage up from calculation, exploiting another comparative advantage of computers: path search underpinned 'Deep Blue's' success.[1] The second assumes that simplification is merited, and so tries to deliver a comprehensible final outcome. The third arises because empirical modeling problems can daunt even experts as unmodeled non-stationarities can seriously distort outcomes. The fourth invokes a learning step, because a good algorithm can incorporate new developments, as *Autometrics* does over (say) Hoover and Perez (1999). The fifth could even apply from different starting general models when all the extra variables were actually irrelevant.

Nevertheless, the best model selection approaches cannot be expected to select the LDGP on every occasion, even when *Gets* is directly applicable and the initial general unrestricted model (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. Since models are often constructed with a specific purpose in mind, they 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. Thus, there are many ways to judge the 'success' of selection algorithms, including:

[1] In May 1997, Deep Blue was the first computer ever to beat a reigning chess world champion. More information is at `www.research.ibm.com/deepblue/home/html/b.html`.

(A) High frequency of recovery of the LDGP;

(B) improved inference about parameters of interest over the GUM;

(C) improved forecasting over other selection methods;

(D) working well for 'realistic' LDGPs;

(E) ability to recover the LDGP starting from the GUM similar to when starting from the LDGP;

(F) operating characteristics of the algorithm match its theory;

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

The first is overly demanding, as it may be nearly impossible to find the LDGP even when commencing from it (e.g., some of the variables may have $|\mathsf{t}| < 0.1$). The second 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 (C) is the objective, including using the GUM, other selection methods, averages over a class of model, factor methods, robust devices, neural nets. However, in processes subject to breaks, in-sample performance need not be a reliable guide to later forecasting success. There are also many possible contenders for (D), 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), albeit with different properties in different states of nature. For (E), a distinction must be made between costs of inference and costs of search. The former are inevitable when tests have non-zero null, and non-unit alternative, rejection frequencies, whereas the latter are additional to commencing from the LDGP. Only the costs of search are really due to selecting, as the costs of inference would confront any investigator who began from the LDGP, but could not be certain that the specification was indeed correct — and omniscience is not realistic in empirical economics. Operating characteristics for (F) could include that the nominal null rejection frequency matches the actual; that retained parameters of interest are unbiasedly estimated; that MSEs are small, etc. Finally, there is the 'internal criterion' (G) that the algorithm could not do better for the given sample, in that no other model dominates that selected.

We use (E), (F) and (G) below as the main bases for evaluation, noting that all three could in principle be achieved together.

4 Why Gets model selection succeeds

When all the regressors are mutually orthogonal, it is easy to explain why Gets model selection can be highly successful. Consider the perfectly orthogonal regression model:

$$y_t = \sum_{i=1}^{N} \beta_i z_{i,t} + \epsilon_t \tag{1}$$

where $\mathsf{E}[z_{i,t} z_{j,t}] = \lambda_i \delta_{i,j}\ \forall i, j$, where $\delta_{i,j} = 1$ if $i = j$ and zero otherwise, $\epsilon_t \sim \mathsf{IN}[0, \sigma_\epsilon^2]$ and $T >> N$.

After unrestricted estimation of (1), order the N sample t^2-statistics testing H_0: $\beta_j = 0$ 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 for significance level c_α is given by:

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

Variables with large t^2 values are retained and all other variables are then eliminated. Only one decision is needed to implement (3), even for $N = 1000$, and 'repeated testing' clearly 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, of course, feasible, and 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', and should also 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 step-wise 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: also 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 n through retention by $\mathsf{t}^2_{(n)} \geq c_\alpha$. Generalization to instrumental variables estimators is straightforward (see Hendry and Krolzig, 2005), and likelihood estimation in general is feasible (Doornik, 2009).

4.1 Selection effects and bias correction

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

Let $\sigma_{\widehat{\beta}} = \mathsf{E}[\widehat{\sigma}_{\widehat{\beta}}]$ 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):

$$\psi^* = \mathsf{E}\left[\mathsf{t}_{\widetilde{\beta}} \mid \left|\mathsf{t}_{\widetilde{\beta}}\right| > c_\alpha; \psi\right] = \psi + \frac{\phi\,(c_\alpha - \psi) - \phi\,(-c_\alpha - \psi)}{1 - \Phi\,(c_\alpha - \psi) + \Phi\,(-c_\alpha - \psi)} = \psi + r\,(\psi, c_\alpha)\,. \tag{4}$$

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

$$\mathsf{E}\left[\widetilde{\beta} \mid \widetilde{\beta} \geq \sigma_{\widetilde{\beta}} c_\alpha\right] = \beta + \sigma_{\widetilde{\beta}} r\,(\psi, c_\alpha) = \beta\left(1 + \psi^{-1} r\,(\psi, c_\alpha)\right), \tag{5}$$

so an unbiased estimator after selection is:

$$\overline{\beta} = \widetilde{\beta}\left(\frac{\psi}{\psi + r\,(\psi, c_\alpha)}\right). \tag{6}$$

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 (4):

$$\psi = \psi^* - r(\psi, c_\alpha). \tag{7}$$

As a first step, replace $r(\psi, c_\alpha)$ in (7) by $r(\mathsf{t}_{\widetilde{\beta}}, c_\alpha)$, and ψ^* by $\mathsf{t}_{\widetilde{\beta}}$:

$$\overline{\mathsf{t}}_{\widetilde{\beta}} = \mathsf{t}_{\widetilde{\beta}} - r\left(\mathsf{t}_{\widetilde{\beta}}, c_\alpha\right). \tag{8}$$

This gives the 1-step bias correction:

$$\overline{\beta} = \widetilde{\beta}\left(\frac{\overline{\mathsf{t}}_{\widetilde{\beta}}}{\mathsf{t}_{\widetilde{\beta}}}\right). \tag{9}$$

Next:

$$\overline{\overline{\mathsf{t}}}_{\widetilde{\beta}} = \mathsf{t}_{\widetilde{\beta}} - r\left(\overline{\mathsf{t}}_{\widetilde{\beta}}, c_\alpha\right), \tag{10}$$

then the two-step bias-corrected parameter estimate is:

$$\overline{\overline{\beta}} = \widetilde{\beta}\left(\frac{\overline{\overline{\mathsf{t}}}_{\widetilde{\beta}}}{\mathsf{t}_{\widetilde{\beta}}}\right). \tag{11}$$

Hendry and Krolzig (2005) show that most, but not all, of the selection bias is corrected for relevant retained variables by (11), at the cost of a small increase in the conditional MSEs. However, 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 any retained irrelevant variables, giving a substantive benefit for both their unconditional and conditional distributions. Thus, despite selecting from a large set of potential variables, nearly unbiased estimates of coefficients and equation standard errors 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, but could be for fat-tailed error processes at tighter α, an issue examined in section 6.1.

4.2 Monte Carlo simulation for $N = 1000$

We now illustrate the above theory by simulating selection from 1000 variables. The DGP is given by:

$$\begin{aligned} y_t &= \beta_1 x_{1,t} + \cdots + \beta_{10} x_{10,t} + \epsilon_t, && (12)\\ \mathbf{x}_t &\sim \mathsf{IN}_{1000}\left[\mathbf{0}, \mathbf{\Omega}\right], && (13)\\ \epsilon_t &\sim \mathsf{IN}\left[0, 1\right], && (14) \end{aligned}$$

where $\mathbf{x}_t' = (x_{1,t}, \cdots, x_{1000,t})$. We set $\mathbf{\Omega} = \mathbf{I}_{1000}$ for simplicity, keeping the regressors fixed between experiments, and use $T = 2000$ observations. The DGP coefficients and non-centralities, ψ, are reported in table 1, together with the theoretical powers of t-tests on the individual coefficients.

The GUM contains all 1000 regressors and a constant term:

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

The DGP has the first $n = 10$ variables relevant, so 991 variables are irrelevant in the GUM (including the intercept).

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

	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 1: Coefficients β_i, non-centralities ψ_i, and theoretical retention probabilities.

outcomes for $\alpha = 1\%$ and 0.1% using $M = 1000$ replications. Because the 'size' of a test statistic has a definition which is only precise for a similar test, and the word is anyway ambiguous in many settings (such as sample size), we use the term '*gauge*' to denote the empirical null rejection frequency of selection tests. Similarly, retaining relevant variables by rejecting their null no longer corresponds to the conventional notion of 'power', so we use the term '*potency*' to denote the average non-null rejection frequency of such tests.

The potencies and gauges are recorded in table 2. The gauges are not significantly different from their nominal sizes α, so the selection is not 'oversized', and the potencies do not deviate relative to 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 variables.

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

Table 2: Potency and gauge for 1-cut selection with 1000 variables.

In addition to gauges and potencies, we report MSEs after model selection. Let $\widehat{\beta}_{k,i}$ denote the OLS estimate of the coefficient on x_{kt} in the GUM for replication i. Let $\widetilde{\beta}_{k,i}$ be the OLS estimate after model selection, so $\widetilde{\beta}_{k,i} = 0$ when x_{kt} was not selected in the final model. We calculate the following MSEs, where $1(\cdot)$ is the indicator variable:

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

The unconditional MSE (denoted UMSE) substitutes zeros when a variable is not selected. The conditional MSE (CMSE) is computed over the retained variables only.

In DGP (12), $\beta_1 = \cdots = \beta_n \neq 0$ (with $n = 10$) and $\beta_0 = \beta_{n+1} = \cdots = \beta_N = 0$, so:

$$\begin{aligned}
\text{retention rate } \widetilde{\mathsf{p}}_k &= \tfrac{1}{M}\textstyle\sum_{i=1}^{M}1(\widetilde{\beta}_{k,i}\neq 0), \quad k=1,\ldots,N, \\
\text{potency} &= \tfrac{1}{n}\textstyle\sum_{k=1}^{n}\widetilde{\mathsf{p}}_k, \\
\text{gauge} &= \tfrac{1}{N-n+1}\left(\widetilde{\mathsf{p}}_0+\textstyle\sum_{k=n+1}^{N}\widetilde{\mathsf{p}}_k\right).
\end{aligned}$$

Figure 1 shows that the retention rates for individual relevant variables are as expected from the theory. The CMSEs are always below the UMSEs for the relevant variables (bottom graphs in Fig. 1), with the exception of β_1 at 0.1%.

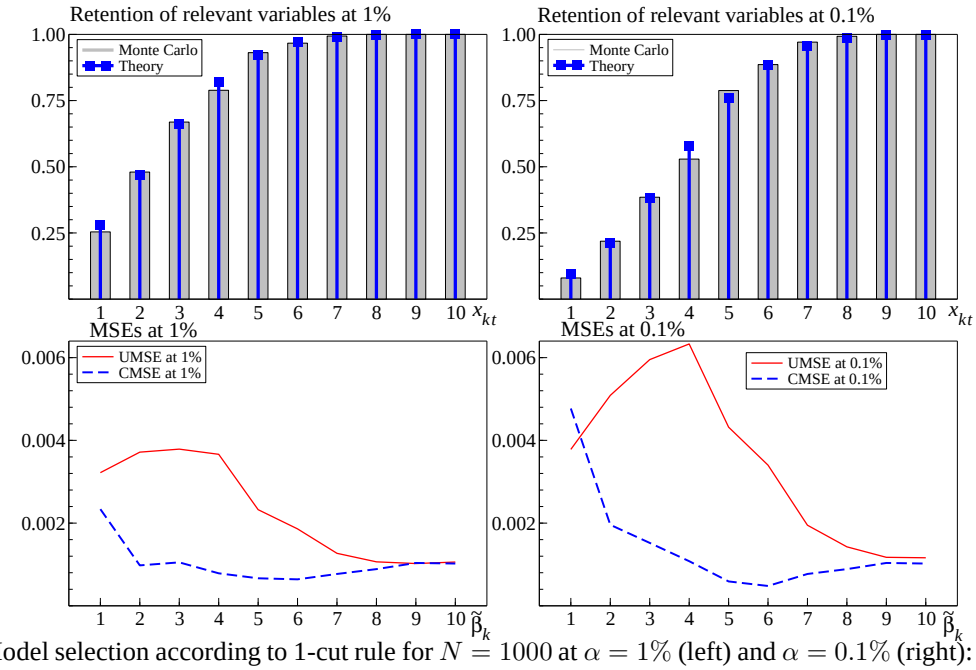

Figure 1: Model selection according to 1-cut rule for $N = 1000$ at $\alpha = 1\%$ (left) and $\alpha = 0.1\%$ (right): retention rates $\widetilde{p}_k$ of relevant variables $x_1, \ldots, x_{10}$ (top graphs), UMSE_k and CMSE_k (bottom graphs).

4.3 Impact of bias correction on MSEs

In 1-cut selection, all variables are significant at c_α by design. 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 deletion makes 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 that are less than the critical value are in a sense irrelevant, and the bias correction formula is non-linear at the critical value, we apply it only to significant retained variables, setting insignificant variables to zero. The result is a substantial improvement in MSE for the irrelevant variables as it downweights chance significance. The 2-step correction in (11) is preferable to the 1-step correction for irrelevant variables, but there is a slight increase in MSE for relevant variables. The preferred correction depends on the non-centralities of the relevant variables and the number of irrelevant variables.

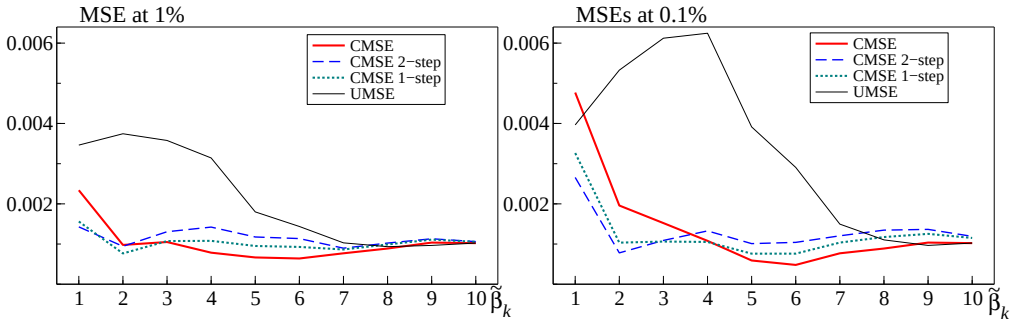

Figure 2: Impact of bias correction on CMSE_k for relevant variables ($\alpha = 1\%$ left and $\alpha = 0.1\%$ right).

Figure 2 records the MSEs of the bias-corrected relevant coefficient estimates in their conditional distributions. Here, the impact of bias correction is quite small. Table 3 shows that the bias corrections for the retained irrelevant variables substantially reduce MSEs.

α	1%	0.1%	1%	0.1%
	average CMSE over 990 irrelevant variables		average CMSE over 10 relevant variables	
using uncorrected $\widetilde{\beta}$	0.84%	1.23%	0.10%	0.14%
using $\overline{\beta}$ after 1-step bias correction	0.53%	0.82%	0.10%	0.13%
using $\overline{\overline{\beta}}$ after 2-step bias correction	0.38%	0.60%	0.12%	0.13%

Table 3: Average CMSE of selected relevant variables, and average CMSE of selected irrelevant variables (excluding β_0), with and without bias correction, $M = 1000$.

Only 2-step bias corrections are reported in the remainder. To record more detail about selection outcomes, we next consider an experiment with a much smaller number of candidate regressors based on a design formulated by Qin and Reed (2008).

5 Comparisons with *Autometrics*

The experimental design is now given by $N = 10$ and $T = 75$:

$$y_t = \beta_0 + \beta_1 x_{1,t} + \cdots + \beta_{10} x_{10,t} + \epsilon_t, \tag{15}$$

$$\mathbf{x}_t \sim \mathsf{IN}_{10}\left[\mathbf{0}, \mathbf{I}_{10}\right], \tag{16}$$

$$\epsilon_t \sim \mathsf{IN}\left[0, \left(0.4 \times n^{0.5}\right)^2\right], \quad n = 1, \ldots, 10, \ t = 1, \ldots, T \tag{17}$$

where $\mathbf{x}_t' = (x_{1,t}, \cdots, x_{10,t})$. The $\mathbf{x}_t$ are fixed across replications as before. Equations (15)–(17) 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 $\beta_0 = 5$. Table 4 reports the non-centralities, ψ.

n	1	2	3	4	5	6	7	8	9	10
$\psi_1 \ldots \psi_n$	21.6	15.3	12.5	10.8	9.7	8.8	8.2	7.7	7.2	6.9

Table 4: Non-centralities for simulation experiments (15)–(17).

The GUM is the same for all 10 DGPs:

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

5.1 1-cut selection and *Autometrics* comparisons

We first investigate how the general search algorithm in *Autometrics* performs relative to the 1-cut selection rule in terms of (E)-(F) in section 3 above. Their comparative gauges are recorded in figure 3 where *Autometrics* selects either with diagnostic tracking or without (the potency ratio is unity for all significance levels, so is not shown). *Autometrics* in default mode (i.e., with diagnostic tracking switched on) is consistently 'overgauged' by about 1 percentage point (see section 5.2).

Figure 4 records the ratio of MSE of *Autometrics* selection to the 1-cut rule for both unconditional and conditional distributions, with no diagnostic tests and no bias correction, for $M = 1000$. 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 irrelevant variables but there is some

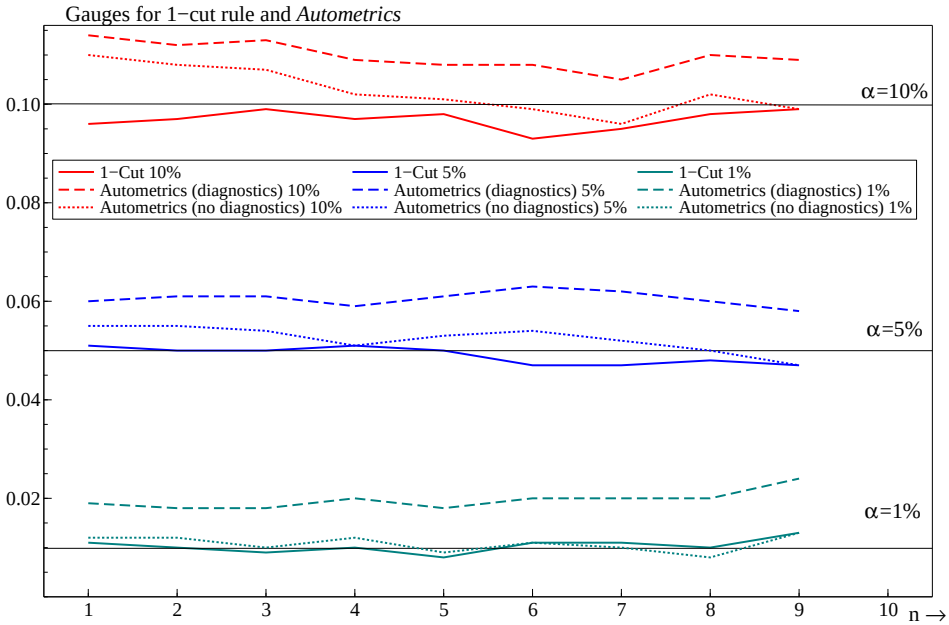

Figure 3: Gauges for 1-cut rule (solid lines), *Autometrics* with diagnostic tracking (dashed lines) and *Autometrics* without diagnostic tracking (dotted lines) for $\alpha = 0.01, 0.05, 0.1$. The horizontal axis represents the $n = 1, \ldots, 10$ DGPs, each with n relevant variables (and $10 - n$ irrelevant).

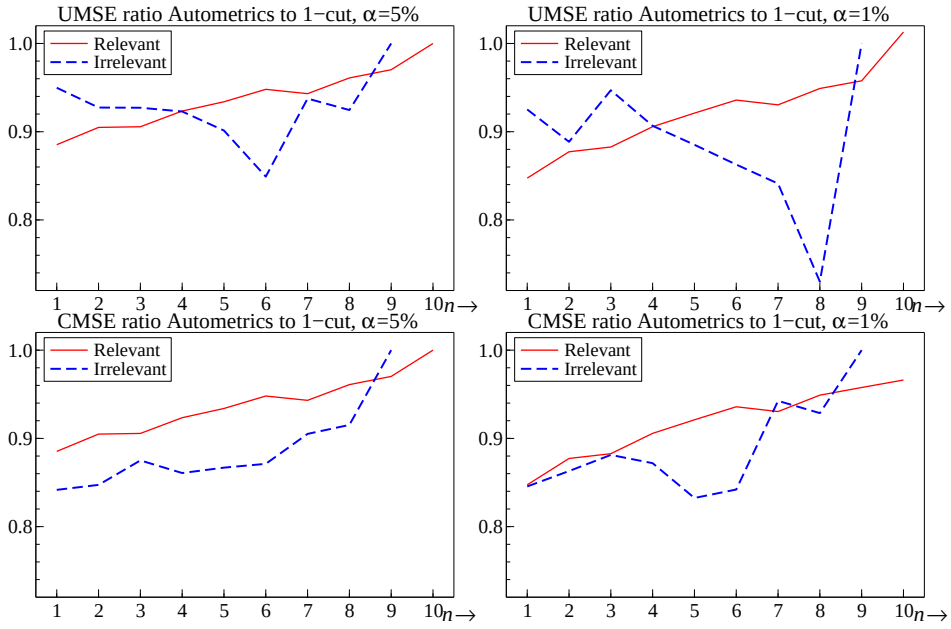

Figure 4: Ratios of MSEs for *Autometrics* to 1-cut rule as n changes

advantage to selection using *Autometrics* for the relevant variables, where the ratios are uniformly less than unity. The benefits to selection are largest when there are few relevant variables that are highly significant. Conditionally, *Autometrics* manages to outperform the 1-cut rule in most cases. Thus, we have established that there is little loss, and perhaps a gain, 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.

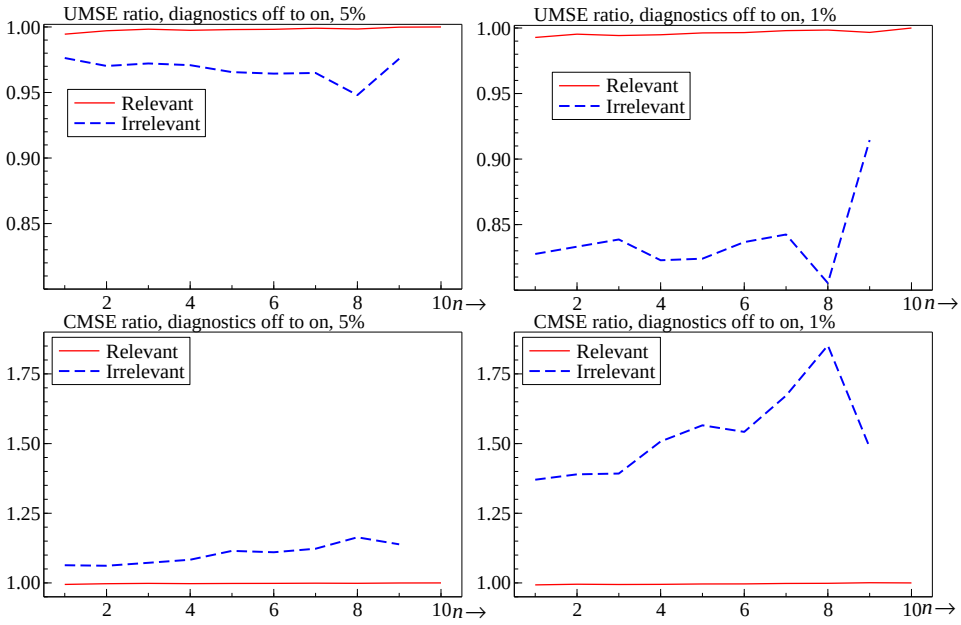

Figure 5: Ratios of MSEs with diagnostic tests off to on for unconditional and conditional distributions

5.2 Impact of diagnostic tests

Figure 3 also compares the gauges for *Autometrics* with diagnostic tracking switched on versus off. 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 essentially equal to the nominal significance level. The difference is due to irrelevant variables proxying part of a chance departure from the null of one of the five mis-specification tests or the encompassing check, and then being retained despite insignificance.

When p_d is the probability of false rejection on each diagnostic test (zero if off), then the impact on the gauge, g, due to an insignificant irrelevant variable being retained to prevent a diagnostic test being significant, is approximately:

$$g \simeq \alpha + \mathsf{p}_d \left(1 + \alpha \left(N - n\right)\right).$$

Excess rejection is relatively largest when $\mathsf{p}_d \geq \alpha$: here, $\mathsf{p}_d = 0.01$, so for $(N - n) < 10$, $g_{0.1} \simeq 0.12$, $g_{0.05} \simeq 0.065$ and $g_{0.01} \simeq 0.02$, which provide a close match to fig. 3.

Figure 5 records the ratio of the UMSEs with diagnostic tests switched off to on in the top panel, and the same for the CMSEs in the bottom panel, averaging within relevant and irrelevant variables. Switching the diagnostics off generally improves the UMSEs, 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 UMSEs, but those irrelevant variables that are retained are now more significant than with the diagnostics on. The impact is largest at tight significance levels: at 10% the ratios are so close to unity they are not plotted.

6 Impulse saturation

Impulse saturation (see Hendry *et al.*, 2008, and Johansen and Nielsen, 2009) includes an indicator for every observation, entered (in the simplest case) in blocks of $T/2$, with the significant outcomes retained. First, add half the indicators, select as usual, record the outcome, then drop that set of impulses and add the other half, selecting again. These first two steps correspond to 'dummying out' $T/2$ observations for

estimation, since impulses are mutually orthogonal. Now combine the significant impulses and select as usual: then αT indicators will be retained on average. Setting $\alpha \leq r/T$ maintains the average false null retention at r 'outliers', equivalent to 'losing' r observations, which is a small efficiency loss for testing the potential relevance of T variables when r is small (e.g., 1). The theory generalizes to more, and unequal, splits, as well as dynamic models. Impulse saturation is under the null of no outliers, but with the aim of detecting and removing outliers and location shifts when they are present.

Autometrics uses its more sophisticated general algorithm even though the impulses are orthogonal, and generally tries several block divisions. Using the same experimental design as before, we record the gauge and potency *averaged across all* $n = 1, \ldots, 10$ experiments in table 5, in which no diagnostic testing is undertaken in *Autometrics*. With no impulse saturation, the gauge is close to the nominal significance level: see fig. 3. With impulse saturation, at our recommended tight significance levels, selection has the appropriate properties. The gauge is slightly too large at a 1% nominal significance level (when 0.75 dummies are retained on average), but slightly too small at the 0.1% significance level (low probability of retaining any dummies). The average potencies are all close to unity reflecting the high non-centralities of the relevant variables.

α	1%		0.1%	
	no impulse saturation	impulse saturation	no impulse saturation	impulse saturation
ave. gauge	1.22%	1.43%	0.12%	0.08%
ave. potency	99.98%	99.97%	99.87%	99.82%

Table 5: Gauge and potency for selection averaged across all $n = 1, \ldots, 10$ experiments with and without impulse saturation. *Autometrics* with no diagnostic testing.

Figure 6 records the ratio of MSEs without saturation to impulse saturation. Under the null, applying impulse saturation at tight significance levels the costs are small, although the MSEs of the irrelevant variables in particular are larger than without impulse saturation at 1%. Correlations between dummies and retained irrelevant variables could increase the weight of the retained irrelevant variables leading to an increase in MSE. At very tight significance levels (0.1%) so few dummies are retained that it has little impact on the MSE, although in some cases impulse saturation could even improve the MSE under the null that none are significant.

6.1 Impulse saturation in fat-tailed distributions

Impulse saturation is designed to detect outliers and location shifts, but we also assess its impact for a fat-tailed distribution, focusing on the student-t distribution with 3 degrees of freedom. The design of the experiment is identical to (15) and (16), but (17) becomes:

$$\epsilon_t \sim \left(0.4 \times n^{0.5}\right) \times \mathsf{t}_3, \quad n = 1, \ldots, 10 \tag{18}$$

Autometrics checks normality in its batch of diagnostic tests. If it rejects, the p-value of the test is reduced, but the program tries to return to the original p-value at a later stage in selection, and may retain irrelevant variables to ensure the diagnostic test is passed.

Table 6 records the average gauge and potency across all $n = 1, \ldots, 10$ experiments using t_3, with diagnostic testing, without, and with impulse saturation. If diagnostic testing is applied, and the DGP is incorrectly assumed to be normal, the gauge is higher than the nominal significance level, and is much higher at tight significance levels (7% for $\alpha = 0.1\%$). Omitting diagnostic testing improves the gauge, but it is still too large. While the critical values used in the selection algorithm will be incorrect for a t_3-distribution, this does not have a substantial impact on the gauge as the regressors are

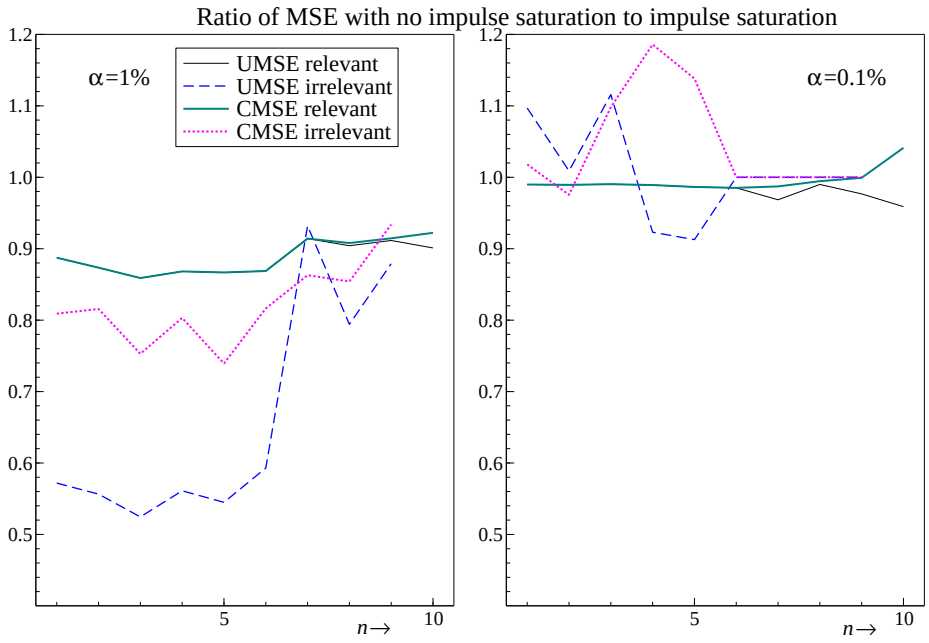

Figure 6: Ratios of MSEs without impulse saturation to impulse saturation

normally distributed. Applying impulse saturation reduces the gauge relative to no impulse saturation and diagnostic testing, but the definition of gauge is ambiguous. Retained dummies are counted as irrelevant variables since they do not enter the DGP, and therefore contribute to gauge, but the fat-tailed distribution implies that some of these are indeed extreme observations and should be retained, so could be considered as contributing to potency. Thus, there is no objective measure of the success of impulse saturation in the form of potency and gauge. Instead, we calculate the average retention probability of the retained irrelevant variables not counting dummies. The gauge is substantially reduced in all cases, but is still larger than the nominal significance level, even for a 0.1% significance level.

impulse saturation	no	no	yes	no	no	yes
diagnostic tracking	yes	no	no	yes	no	no
α		1%			0.1%	
ave. gauge	8.62%	1.37%	4.91%	6.80%	0.52%	1.37%
ave. gauge (exc. dummies)	–	–	1.86%	–	–	0.23%
ave. potency	96.26%	96.30%	98.78%	92.42%	92.45%	91.40%

Table 6: Gauge and potency for t_3-distribution. Average across all $n = 1, \ldots, 10$ experiments with and without impulse saturation and diagnostics.

Figure 7 records the ratio of MSEs without saturation to impulse saturation for t_3. The benefit of impulse saturation is observed by smaller MSEs for the coefficient estimates of the retained regressors at significance levels of 1% or 0.1%. Although the gauge is larger than for selection without impulse saturation and no diagnostic testing, the coefficient estimates for the retained variables are much closer to their DGP values as the dummies account for the fat-tails, bringing the distribution conditional on the dummies closer to a normal distribution. To check the impact of impulse saturation, figure 8 compares the conditional distributions for $\tilde{\beta}_0, \ldots, \tilde{\beta}_{10}$ where $n = 5$, such that $\beta_0 = 5$, $\beta_1, \ldots, \beta_5 = 1$ and $\beta_6, \ldots, \beta_{10} = 0$. The impulse saturated distributions are super-imposed on the conditional distributions. For the relevant variables, the distributions are similar, although the distributions without impulse saturation have slightly fatter tails. The long tails for the non-saturated conditional distributions of the

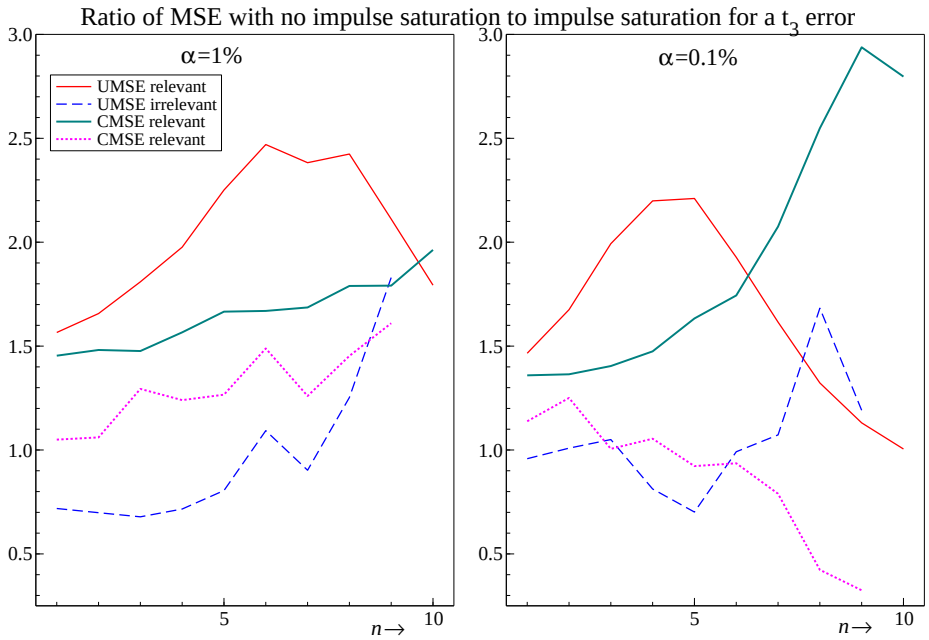

Figure 7: Ratios of MSEs without impulse saturation to impulse saturation for a t_3-distribution with no diagnostic testing.

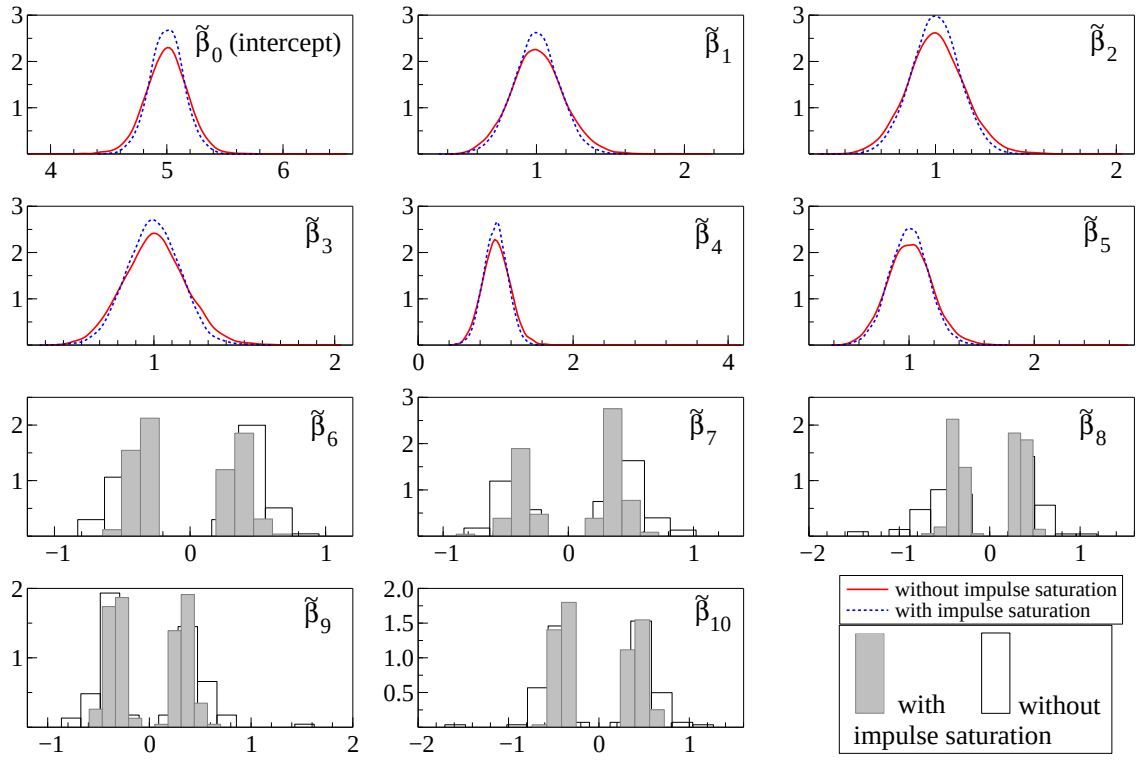

Figure 8: Conditional distributions of estimates with (dark) and without (light) impulse saturation for the $n = 5$ experiment with a t_3 error distribution at $\alpha = 1\%$, $M = 10\,000$.

coefficient estimates for the irrelevant variables relative to the saturated outcomes are evident. Thus, if irrelevant variables are retained, impulse saturation in the presence of fat-tails will imply the reported coefficient estimates are smaller than would otherwise be reported and bias correction will downweight even further, providing insurance against highly significant but irrelevant variables.

7 Selecting regressions when there are multiple breaks

Having established that testing T impulses need not induce a large efficiency loss when they are irrelevant, we now apply impulse saturation to detect outliers and locations shifts when they do in fact occur. The generality of impulse saturation allows it to detect many shifts, and we consider up to 20 in 100 observations, only one, and several intermediate settings. Moreover, the shifts can be right at the start or end of the sample, as there is no need to reserve a percentage as with (say) Bai and Perron (1998).

7.1 Impulse saturation for breaks in the mean of a location-scale model

The first set of experiments examine the detectability of various forms of location shift in the simplest setting, but for a range of magnitudes, forms and timing of breaks. The DGPs considered all have $T = 100$ using $M = 1000$ replications to investigate:

- DGP:Bc–a single break in the mean (starting at $T = 81$);
- DGP:B20–20 breaks in the mean (starting at $T = 1$, equally spread);
- DGP:MBc–multiple breaks in the mean (5 breaks of length four, equally spread);
- DGP:Bct–as DGP:Bc, but with a trend in both the DGP and GUM;
- DGP:MBct–as DGP:MBc, but with a trend in both the DGP and GUM;
- DGP:Tc–a break in the trend (trending from $T = 81$, with no trend before);
- DGP:BL–a break in the mean in a stationary autoregression with a zero mean;
- DGP:BLc–a break in the mean in a stationary autoregression with a non-zero mean.

Thus, we used the following sets of DGPs:

DGP:Bc	y_t	$=$	$\delta + \gamma\left(I_{81} + \cdots + I_{100}\right) + u_t,$
DGP:B20	y_t	$=$	$\delta + \gamma\left(I_1 + I_6 + I_{11} + \cdots + I_{96}\right) + u_t,$
DGP:MBc	y_t	$=$	$\delta + \gamma\left(I_1 + I_2 + I_3 + I_4 + I_{24} + \cdots + I_{27} + I_{49} + \cdots + I_{52} + I_{74} + \cdots + I_{77} + I_{97} + \cdots + I_{100}\right) + u_t,$
DGP:Bct	y_t	$=$	$\delta + \gamma\left(I_{81} + \cdots + I_{100}\right) + 0.02t + u_t,$
DGP:MBct	y_t	$=$	$\delta + \gamma\left(I_1 + I_2 + I_3 + I_4 + I_{24} + \cdots + I_{27} + I_{49} + \cdots + I_{52} + I_{74} + \cdots + I_{77} + I_{97} + \cdots + I_{100}\right) + 0.02t + u_t,$
DGP:Tc	y_t	$=$	$\delta + \gamma\left(\frac{1}{20}I_{81} + \frac{2}{20}I_{82} + \cdots + \frac{20}{20}I_{100}\right) + u_t.$
DGP:BL	y_t	$=$	$\gamma\left(I_{81} + \cdots + I_{100}\right) + 0.5y_{t-1} + u_t, \quad y_0 = 0,$
DGP:BLc	y_t	$=$	$2 + \gamma\left(I_{81} + \cdots + I_{100}\right) + 0.5y_{t-1} + u_t, \quad y_0 = 0,$
	u_t	$\sim$	$\mathsf{IN}[0, 1]; \quad \delta = 0, 1$ as noted below

The *Autometrics* algorithm was started from the following GUMs, where 'fixed' entails that the relevant variable cannot be eliminated by selection:

GUM:Ic	y_t on constant (free) and T dummies for DGP:Bc, DGP:B20, DGP:MBc, DGP:Tc;
GUM:Ict	y_t on constant (free), T dummies, and trend for DGP:Bct, DGP:MBct;
GUM:Ic	y_t on constant (fixed and free), T dummies, and y_{t-1} for DGP:BL, DGP:BcL.

We first report in detail the results for DGP:Bc and DGP:B20 when $\delta = 0$ for $\gamma = 1$ up to $\gamma = 5$.

It is clearly much easier to detect a single break of length 20 than twenty breaks of 1 period when γ is small, but the potencies rapidly rise towards unity in both cases as γ grows to 5. While 20 'shifts' in a sample of 100 is unlikely in practice, the ability to find them in such a contaminated case is encouraging.

	DGP:Bc					
$\alpha = 1\%$	$\gamma = 0$	$\gamma = 1$	$\gamma = 2$	$\gamma = 3$	$\gamma = 4$	$\gamma = 5$
Gauge %	1.5	1.2	0.9	0.3	0.7	1.1
Potency %	—	4.6	25.6	52.6	86.3	99.0
	DGP:B20					
$\alpha = 1\%$	$\gamma = 0$	$\gamma = 1$	$\gamma = 2$	$\gamma = 3$	$\gamma = 4$	$\gamma = 5$
Gauge %	1.5	1.0	0.4	0.3	1.0	0.8
Potency %	—	3.5	7.9	24.2	67.1	90.2

Table 7: *Autometrics* model selection for impulse saturation in location-scale DGPs with breaks at the end and multiple breaks.

	$\gamma = 3$	$\gamma = 4$	$\gamma = 5$	$\gamma = 3$	$\gamma = 4$	$\gamma = 5$
	Autometrics, constant fixed, $\alpha = 1\%$					
	DGP:Bc			DGP:Bct		
Gauge %	0.4	0.7	1.1	4.7	1.6	1.1
Potency %	55.2	87.1	99.1	29.4	68.9	92.2
	DGP:MBc			DGP:MBct		
Gauge %	0.4	0.7	1.0	0.5	0.8	1.1
Potency %	38.8	78.4	96.5	42.2	77.8	94.9
	Step-wise regression, constant fixed, $\alpha = 1\%$					
	DGP:Bc			DGP:Bct		
Gauge %	0.1	0.0	0.1	0.7	0.4	0.2
Potency %	13.8	16.2	16.5	6.2	6.2	5.9
	DGP:MBc			DGP:MBct		
Gauge %	0.1	0.0	0.1	0.1	0.0	0.2
Potency %	14.2	16.9	18.2	18.0	20.2	24.6

Table 8: *Autometrics* and step-wise model selection for impulse saturation in location-scale (and trend) DGPs with breaks at the end and multiple breaks. All estimated models have a constant term.

Next, *Autometrics* model selection was also compared to step-wise regression to highlight the improvements from path search, now with $\delta = 1$ (we report DGP:Bc as a comparator with table 7). The potency of *Autometrics* rises rapidly with γ in all four DGPs, with almost all breaks detected by $\gamma = 5$. The setting with five breaks of length 4 is realistic, and has a potency between that of the two cases in table 7. *Autometrics* highly outperforms step-wise, a pattern that occurred in every experiment, so we do not report the latter henceforth (step-wise results are available on request).

7.2 Impulse saturation for breaks in the mean of a stationary autoregression

Table 9 reports the simulation results. *Autometrics* selects the wrong model for DGP:BLc when the constant is free, namely a unit-root model without a constant and with some additional dummies.

The treatment of the constant matters greatly for the *Autometrics* results as follows:

- Constant is free: when the constant is free, it is not selected. This is fine for DGP:BL, but not for DGP:BLc. In most cases, the model selected for DGP:BLc has one (or two) dummies at the start of the break, and a unit root.

- Constant restricted: when a constant is forced to enter all models, *Autometrics* has high power to

	$\gamma = 5$	$\gamma = 8$	$\gamma = 10$	$\gamma = 5$	$\gamma = 8$	$\gamma = 10$
	Autometrics, constant free, $\alpha = 1\%$					
	DGP:BL			DGP:BLc		
Gauge %	1.2	1.2	1.2	1.5	1.5	1.5
Potency %	44.2	80.9	91.2	12.7	15.6	16.7
	Autometrics, constant fixed, $\alpha = 1\%$					
	DGP:BL			DGP:BLc		
Gauge %	1.2	1.2	1.2	1.2	1.2	1.2
Potency %	48.5	83.5	92.7	49.3	84.7	93.5

Table 9: *Autometrics* results for impulse saturation using location-scale DGPs with breaks at end and multiple breaks.

detect the correct model for both DGPs.

7.3 Impulse saturation in unit-root models

An impulse in a unit-root model entails a step shift in the level of the series, so is the most realistic alternative in this setting. Thus, we consider the following sets of experiments.

DGP:IUc	y_t	$=$	$0.2 + \gamma I_{81} + y_{t-1} + u_t,$
DGP:BUc	y_t	$=$	$0.2 + \gamma\left(I_{81} + \cdots + I_{100}\right) + y_{t-1} + u_t,$
DGP:MIUc	y_t	$=$	$0.2 + \gamma\left(I_1 + I_{24} + I_{49} + I_{74} + I_{97}\right) + y_{t-1} + u_t,$
DGP:MBUc	y_t	$=$	$0.2 + \gamma\left(I_1 + \cdots + I_4 + I_{24} + \cdots + I_{27} + I_{49} + \cdots + I_{52} + I_{74} + \cdots + I_{77} + I_{97} + \cdots + I_{100}\right) + y_{t-1} + u_t,$
			$u_t \sim \mathsf{IN}[0,1],\ y_{-100} = 0,\ t = -99, \ldots, 0, 1, \ldots, 100$

The corresponding initial general models are:

GUM:ILct	y_t on constant and T dummies, y_{t-1}, and trend, $t = 1, \ldots, 100$;
DGUM:ILct	Δy_t on constant, T dummies, y_{t-1}, and trend, $t = 1, \ldots, 100$.

DGUM has Δy_t as the dependent variable, while the regressors include y_{t-1}. After model selection, the final model is re-estimated with y_t as the dependent variable and adding y_{t-1} as a regressor (if necessary). Table 10 shows that gauge remains well controlled at the nominal significance level, and potency to detect the breaks is reasonably high but rises slowly with their magnitude. Also, the transformation to DGUM varies between moderately and strongly beneficial, probably because it allows y_{t-1} to be eliminated more often than with the levels.

7.4 Level saturation in autoregressions with and without unit roots

In level saturation, all impulse and step dummies are used:

1. $I_s = 0$ for period s, 0 otherwise,
2. $S_s = 1$ for periods $t \geq s$, 0 otherwise,

Now the basic GUM is:

$$y_t = \mu + \sum_{s=1}^{T} \delta_{1,s} I_s + \sum_{s=2}^{T} \delta_{2,s} S_s + \epsilon_t, \quad \epsilon_t \sim \mathsf{IN}[0, \sigma_\epsilon^2], \quad t = 1, \ldots, T.$$

	GUM			DGUM		
	$\gamma = 3$	$\gamma = 4$	$\gamma = 5$	$\gamma = 3$	$\gamma = 4$	$\gamma = 5$
	Autometrics, $\alpha = 1\%$, constant free					
	DGP:IUc			DGP:IUc		
Gauge %	1.5	1.4	1.4	1.7	1.7	1.7
Potency %	60.8	69.4	71.5	65.8	73.0	74.6
	DGP:BUc			DGP:BUc		
Gauge %	1.0	1.0	1.1	1.2	1.3	1.4
Potency %	22.5	39.3	56.9	28.2	65.9	87.9
	DGP:MIUc			DGP:MIUc		
Gauge %	1.1	1.2	1.2	1.6	1.7	1.7
Potency %	55.9	74.8	83.7	64.8	82.4	87.8
	DGP:MBUc			DGP:MBUc		
Gauge %	1.2	2.6	1.9	1.0	1.1	1.2
Potency %	33.2	51.6	53.7	44.0	75.6	92.0

Table 10: *Autometrics* results for impulse saturation using a unit-root DGP with breaks at end and multiple breaks.

The DGP is rewritten in terms of S_s, which affects counting for gauge and potency:

DGP:BLc	$y_t = 2 + \gamma S_{81} + 0.5 y_{t-1} + u_t, \quad y_0 = 0, t = 1, \ldots, 100,$
	$u_t \sim \mathsf{IN}[0, 1]$
GUM:SLc	y_t on constant, T dummies, $T - 1$ levels, y_{t-1}, $t = 1, \ldots, 100$.
DGUM:SLc	Δy_t on constant, T dummies, $T - 1$ levels, y_{t-1}, $t = 1, \ldots, 100$.
DGP:BUc	$y_t = 0.2 + \gamma S_{81} + y_{t-1} + u_t, y_{-100} = 0, t = -99, \ldots, 0, 1, \ldots, 100,$
	$u_t \sim \mathsf{IN}[0, 1]$
GUM:SLct	y_t on constant, T dummies, $T - 1$ levels, y_{t-1}, and trend, $t = 1, \ldots, 100$.
DGUM:SLct	Δy_t on constant, T dummies, $T - 1$ levels, y_{t-1}, and trend, $t = 1, \ldots, 100$.

With impulse saturation, detecting all the dummies and lagged y, but missing the constant would give a potency of 95%. With level saturation, a similar result would be to detect y_{t-1} and L_{81} for a potency of 66%. Similarly, a gauge of 1% was about 0.8 irrelevant variable before, but will be around 2 in the level saturation.

It is possible that selection would choose some of the correct dummies instead of the level, or just the wrong level with one extra dummy, increasing gauge and decreasing potency. One way around this would be to translate the final model into dummies, and use that to measure gauge and potency. Another approach is to do some post-selection processing. After selection, all variables (with a few exceptions) are significant. We distinguish the following cases:

1. Two step dummies that almost overlap. With the following substitutions:

$$\begin{aligned} S_{t-1}, S_t &\Rightarrow I_{t-1}, S_t, \\ S_{t-2}, S_t &\Rightarrow I_{t-2}, I_{t-1}, S_t, \end{aligned}$$

   it is possible that the step dummy S_t becomes redundant (when the coefficients on the step dummies have opposite signs, but are otherwise not statistically different). If not, the coefficient of S_t will be significantly different from the impulse dummies, so the formulation with impulses is preferred.

2. Overlapping impulse and step dummy can be rewritten as follows:

$$I_{t-1}, S_{t-1} \quad \Rightarrow \quad I_{t-1}, S_t.$$

This will result in the impulse dummy I_{t-1} becoming redundant when the coefficients on I_{t-1} and S_{t-1} have opposite signs, but are otherwise not statistically different. The transformation is only applied when the coefficients have the required opposite signs.

3. Impulse dummy preceding step dummy:

$$\begin{array}{lcl} I_{t-1}, S_t & \Rightarrow & I_{t-1}, S_{t-1}, \\ I_{t-s}, \ldots, I_{t-1}, S_t & \Rightarrow & I_{t-s}, \ldots, I_{t-1}, S_{t-s}. \end{array}$$

If all the impulse dummies preceding the step dummy have the same sign as the step dummy, it is attractive to extend the step dummy backward. This would potentially allow some impulses to disappear.

	Autometrics, Constant free, $\alpha = 1\%$, $M = 1000$					
	GUM			DGUM		
	without post-processing					
	DGP:BLc			DGP:BLc		
	$\gamma = 5$	$\gamma = 8$	$\gamma = 10$	$\gamma = 5$	$\gamma = 8$	$\gamma = 10$
Gauge %	1.7	1.2	1.2	1.5	1.3	1.3
Potency %	64.6	74.7	74.7	83.9	92.8	95.4
	DGP:BUc			DGP:BUc		
	$\gamma = 3$	$\gamma = 4$	$\gamma = 5$	$\gamma = 3$	$\gamma = 4$	$\gamma = 5$
Gauge %	1.0	1.0	1.6	1.4	1.2	1.2
Potency %	53.9	56.0	54.8	62.5	65.6	66.2
	with post-processing					
	DGP:BLc			DGP:BLc		
	$\gamma = 5$	$\gamma = 8$	$\gamma = 10$	$\gamma = 5$	$\gamma = 8$	$\gamma = 10$
Gauge %	1.2	1.1	1.1	1.2	1.1	1.2
Potency %	70.1	73.6	73.8	84.1	92.2	94.4
	DGP:BUc			DGP:BUc		
	$\gamma = 3$	$\gamma = 4$	$\gamma = 5$	$\gamma = 3$	$\gamma = 4$	$\gamma = 5$
Gauge %	0.9	0.9	0.9	1.1	1.0	0.9
Potency %	56.9	61.4	63.7	62.7	66.3	67.8

Table 11: *Autometrics* results for level saturation using autoregressive DGPs with breaks at end, with and without post-processing.

The top half of Table 11 records the result without post-processing, and the bottom half shows the result with post-processing. The clean-up helps *Autometrics* considerably in accounting for gauge and potency. In these tables, a potency around 30% corresponds to y_{t-1} being almost always found, but not S_{81} or the constant. For a potency of around 60% it is usual that y_{t-1} and S_{81} were found, but not the constant.

7.5 Autoregression with regressors

In the following experiments, the regressors in the DGP, denoted X, are specified as in (19) where β is varied across experiments, but ρ is kept fixed at 0.9 with $x_{i,0} = 0$, and $\beta^* = (T[1-\rho])^{-1/2}\beta$:

$$\mathbf{x}_t'\boldsymbol{\beta} = \sum_{i=1}^{4} \beta^* (x_{i,t} - x_{i,t-1}) \tag{19}$$

$$\begin{array}{lcl} x_{i,t} & = & \rho x_{i,t-1} + v_{i,t}, \;\; i = 1, \ldots, 10; \\ v_{i,t} & \sim & \mathsf{IN}\left[0, (1-\rho)^2\right]; \\ \beta & = & \{2.4, 2.8, \ldots, 4.0\}. \end{array}$$

The regressors added to the GUM are as follows:

$$X = \sum_{i=1}^{10} (\gamma_i x_{i,t} + \delta_i x_{i,t-1})$$

so there are 8 relevant regressors (four at lag zero and four at lag one), and 12 irrelevant ones.

DGP:Lcx	$y_t = 2 + 0.5y_{t-1} + \mathbf{x}_t'\boldsymbol{\beta} + u_t, \quad y_0 = 0, t = 1, \ldots, 100,$
GUM:Lcx	y_t on constant, y_{t-1}, X, $t = 1, \ldots, 100$.
DGP:BLcx	$y_t = 2 + \gamma(I_{81} + \cdots + I_{100}) + 0.5y_{t-1} + \mathbf{x}_t'\boldsymbol{\beta} + u_t, \quad y_0 = 0, t = 1, \ldots, 100,$
GUM:ILcx	y_t on constant, T dummies, y_{t-1}, X, $t = 1, \ldots, 100$.
DGP:BLcx	$y_t = 2 + \gamma S_{81} + 0.5y_{t-1} + \mathbf{x}_t'\boldsymbol{\beta} + u_t, \quad y_0 = 0, t = 1, \ldots, 100,$
GUM:SLcx	y_t on constant, T dummies, $T-1$ levels, y_{t-1}, X, $t = 1, \ldots, 100$.
DGP:Ucx	$y_t = 0.2 + y_{t-1} + \mathbf{x}_t'\boldsymbol{\beta} + u_t, y_{-100} = 0, t = -99, \ldots, 0, 1, \ldots, 100,$
GUM:Lcx	y_t on constant, y_{t-1}, $t = 1, \ldots, 100$.
GUM:Lcxt	y_t on constant, y_{t-1}, and trend, $t = 1, \ldots, 100$.
DGP:BUcx	$y_t = 0.2 + \gamma(I_{81} + \cdots + I_{100}) + y_{t-1} + \mathbf{x}_t'\boldsymbol{\beta} + u_t, y_{-100} = 0, t = -99, \ldots, 0, 1, \ldots, 100,$
GUM:ILcxt	y_t on constant, T dummies, and trend, y_{t-1}, X, $t = 1, \ldots, 100$.
DGP:BUcx	$y_t = 0.2 + \gamma S_{81} + y_{t-1} + \mathbf{x}_t'\boldsymbol{\beta} + u_t, y_{-100} = 0, t = -99, \ldots, 0, 1, \ldots, 100,$
GUM:SLcxt	y_t on constant, T dummies, and trend, $T-1$ levels, y_{t-1}, X, $t = 1, \ldots, 100$.

The gauge for *Autometrics* in DGP:Ucx, GUM:Lcxt is around 10%, see Table 12, because the trend is almost always included in the final model, even though it is not in the DGP. In that case, the constant may be deleted, which suggests that a GUM with fixed constant and free trend would be better.

Importantly, when there are breaks–as occurs all too often in practice–gauge is well controlled and potency is again moderate to high, especially for the unit root case: see Table 13.

8 Conclusion

Setting the nominal rejection frequency of individual selection tests at $\alpha \leq 1/N \to 0$ as $T \to \infty$, on average one irrelevant variable will be retained as adventitiously significant out of N candidates. Thus, there is little difficulty in eliminating almost all of the irrelevant variables when starting from the GUM (a small cost of search). The so-called overall 'size' of the selection procedure, namely $1-(1-\alpha)^N$, can be large, but is uninformative about the success of a simplification process that on average correctly eliminates $(1-\alpha)N$ irrelevant variables. Despite very large numbers of irrelevant candidate regressors, *Autometrics* has a null rejection frequency (gauge) close to the nominal size, somewhat increased by the

	$\beta = 2.4$	$\beta = 2.8$	$\beta = 3.2$	$\beta = 3.6$	$\beta = 4.0$
	Autometrics, Constant free, $\alpha = 1\%$, $M = 1000$				
	DGP:Lcx, GUM:Lcx				
Gauge %	4.3	4.0	3.2	2.4	2.3
Potency %	61.8	75.2	87.3	94.6	98.0
	DGP:Ucx, GUM:Lcx				
Gauge %	3.7	3.4	2.9	2.8	2.5
Potency %	58.1	70.8	82.8	87.9	91.0
	DGP:Ucx, GUM:Lcxt				
Gauge %	11.3	10.3	10.2	9.3	8.5
Potency %	57.5	69.0	78.8	85.6	89.5

Table 12: *Autometrics* results for stationary and unit-root autoregressive DGPs without breaks.

	$\beta = 2.4$	$\beta = 2.8$	$\beta = 3.2$	$\beta = 3.6$	$\beta = 4.0$
	Autometrics, Constant free, $\alpha = 1\%$, $M = 1000$				
	DGP:BLcx, GUM:ILcx, $\gamma = 10$				
Gauge %	1.6	1.6	1.6	1.7	1.6
Potency %	34.1	38.9	43.4	48.5	53.2
	DGP:BLcx, GUM:SLcx, $\gamma = 10$				
Gauge %	1.8	1.8	1.7	1.5	1.5
Potency %	58.1	70.0	79.6	88.7	93.0
	DGP:BUcx, GUM:ILcxt, $\gamma = 5$				
Gauge %	3.2	3.4	3.3	3.2	3.2
Potency %	32.9	35.8	38.4	40.8	43.5
	DGP:BUcx,GUM:SLcxt, $\gamma = 5$				
Gauge %	1.5	1.6	1.5	1.4	1.3
Potency %	51.2	63.0	74.5	81.5	86.3

Table 13: *Autometrics* results for stationary and unit-root autoregressive DGPs with breaks.

need to undertake diagnostic testing for congruence and encompassing tests against the GUM. Moreover, bias correction for selection is very effective at reducing the MSEs of retained irrelevant variables in both unconditional and conditional distributions, at a small cost in increased MSEs for relevant variables.

Autometrics with impulse saturation performs very well in location-scale, location-scale-trend, and stationary autoregressions (with one exception), whereas step-wise regression performs poorly in all these cases: often it has no power to detect breaks. However, *Autometrics* finds a single break at the end harder to detect when there is a trend in the DGP (in comparison to no trend), although this is not the case when there are multiple breaks. Thus, in comparison to a single break, multiple breaks are harder to detect in a DGP without a trend, but easier in the DGP with a trend. Also, *Autometrics* performs poorly when there is a break in a stationary autoregression with an intercept (DGP:BLc), and the constant is a free variable, albeit that is easily fixed.

Autoregressions with a unit root and single outliers, or a few scattered outliers, are one situation where step-wise regression does as well or slightly better than *Autometrics*, although that again ceases to hold for unit-root process with breaks. When estimating using a GUM with the dependent variable in differences and a free constant (DGP:BL and DGP:BLc), *Autometrics* does badly, selecting a unit-root model with a few outliers rather than a stationary autoregression with a shift, yet does well detecting

breaks when there is a unit-root DGP. Moreover, *Autometrics* does well using the new proposal of level saturation, particularly when estimating in differences, and the problem of detecting the constant in the stationary autoregression disappears (although, in these DGPs, the constant is generally difficult to detect).

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 $\mathsf{O}(1/\sqrt{T})$, especially when highly correlated with other regressors (see Pötscher, 1991, and 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 $T \rightarrow \infty$, so 'most' parameter values are unproblematic. When the LDGP is not nested in the GUM, the selected approximation may use the incorrect choice if that is undominated, but in a progressive research strategy when there are intermittent structural breaks in both relevant and irrelevant variables, such a selection will soon be dominated. 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.

Overall, we conclude that model selection based on *Autometrics* using relatively tight significance levels and bias correction is a successful approach, and allows multiple breaks to be tackled. Even though the approach requires both expanding and contracting searches–as there are more regressors than observations–impulse and level saturation allow many breaks to be detected and 'modeled' by dummies.

References

Bai, J., and Perron, P. (1998). Estimating and testing linear models with multiple structural changes. *Econometrica*, **66**, 47–78.

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.

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. (2007). *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*, **Special issue**, forthcoming.

Doornik, J. A. (2009). Autometrics. In Castle, J., and Shephard, N. (eds.), *The Methodology and Practice of Econometrics*. Forthcoming, Oxford: Oxford University Press.

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

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.

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.

Johansen, S., and Nielsen, B. (2009). An analysis of the indicator saturation estimator as a robust regression estimator. In Castle, J. L., and Shephard, N. (eds.), *The Methodology and Practice of Econometrics*. Forthcoming, Oxford: Oxford University Press.

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.

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.

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.

Qin, X., and Reed, W. R. (2008). A comparison of a large number of model selection criteria. Working paper, Economics Department, University of Canterbury, Christchurch, New Zealand.

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.