

# Model Selection when there are Multiple Breaks

Jennifer L. Castle<sup>†</sup>, Jurgen A. Doornik and David F. Hendry<sup>\*</sup>

<sup>†</sup>Magdalen College and Institute for New Economic Thinking at the  
Oxford Martin School, University of Oxford, UK

<sup>\*</sup>Economics Department and Institute for New Economic Thinking at the  
Oxford Martin School, University of Oxford, UK

## Abstract

We consider model selection facing uncertainty over the choice of variables and the occurrence and timing of multiple location shifts. General-to-simple selection is extended by adding an impulse indicator for every observation to the set of candidate regressors: see Johansen and Nielsen (2009). We apply that approach to a fat-tailed distribution, and to processes with breaks: Monte Carlo experiments show its capability of detecting up to 20 shifts in 100 observations, while jointly selecting variables. An illustration to U.S. real interest rates compares impulse-indicator saturation with the procedure in Bai and Perron (1998).

*JEL classifications:* C52, C22.

**KEYWORDS:** Impulse-indicator saturation; Location shifts; Model selection; *Autometrics*.

## 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'.

---

<sup>\*</sup>Corresponding address: Nuffield College, New Road, OX1 1NF Oxford, United Kingdom. Tel.: +44 (0) 1865 278587.

# 1 Introduction

Our contribution concerns modeling situations that involve specification uncertainty over the choice of which variables, lags, functional forms, etc., are relevant and which are irrelevant, jointly with determining the occurrence and timing of multiple breaks affecting a 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 variables' 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 can be more variables,  $N$ , in total than the number of observations,  $T$ , so all cannot be entered from the outset. To resolve this conundrum, general-to-simple (*Gets*) selection must be extended to have expanding as well as contracting searches.

When only contracting searches are needed, Castle et al. (2011) consider *Gets* for a constant model in orthogonal variables. Only one selection decision is required irrespective of the number of regressors  $N < T$ , which approach they call 1-cut. Thus, although there are  $2^N$  possible models, only one general model needs estimated and just one decision is required (namely, what variables are excluded, and hence what are included), so 'repeated testing' does not occur: compare, e.g., Leamer (1983). In 1-cut, after ranking the  $t^2$ -values of all variables from the largest (denoted  $t^2_{(1)}$ ) to the smallest ( $t^2_{(N)}$ ), the decision to retain  $m$  variables is made by  $t^2_{(m)} \geq c_\alpha^2$  when  $t^2_{(m+1)} < c_\alpha^2$  where  $c_\alpha$  is the t-distribution critical value corresponding to the nominal significance level  $\alpha$ . Retention rates for irrelevant variables (called gauge) are close to  $\alpha$ , for small  $\alpha$  (e.g.,  $\alpha \leq 1/N$ ), and can be controlled. Retention rates for relevant variables (called potency) are close to the theoretical power for a one-off test at  $c_\alpha$ . Nevertheless, even with orthogonal regressors, the 1-cut approach does not uniformly dominate the outcome of the general search algorithm *Autometrics*, an Ox Package implementing automatic selection: see Doornik (2009a,b), and Hendry and Doornik (2009).

In non-orthogonal problems, path searches are required to establish 'genuine relevance', which gives the impression of 'repeated testing'. *Autometrics* uses a tree-search to detect and eliminate statistically-insignificant variables, improving on the multi-path search in Hoover and Perez (1999) and Hendry and Krolzig (2005). Such an algorithm does not become stuck in a single-path sequence, where inadvertently eliminating a relevant variable leads to retaining other variables as proxies (as could happen in stepwise regression). A variable is only removed if the new model is a valid reduction of the initial model, so must encompass the initial general unrestricted model (denoted GUM) at the chosen  $\alpha$  when  $N < T$ : see Hendry et al. (2008b) and Doornik (2008). A path terminates when no remaining variables meet the reduction criterion. At the end, there will be one or more non-rejected (called terminal) models. All such models are congruent, undominated, mutually-encompassing representations: see Hendry and Nielsen (2007) for an extensive discussion of these terms. If necessary, a tie-breaker using an information criterion like BIC—see Schwarz (1978)—can make a unique selection, although all terminal models are reported and can be used in, say, forecast combinations.

Path search is not equivalent to selecting the 'best fitting model' (however penalized) from the  $2^N$  possible models. 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 any desired variables can be retained by 'forcing' as discussed below: see Hendry and Johansen (2010). However the choice of the critical value,  $c_\alpha$ , for retention tests affects both  $R^2$  and the number of variables retained,  $m$ . Generalizations to instrumental variables estimators and likelihood estimation are relatively straightforward: see Hendry and Krolzig (2005), and Doornik (2009a), respectively.

Selection affects the distributional properties of the final model's estimates as compared with estimating the local data generating process (LDGP — the DGP in the space of the variables under analysis): see e.g., Hendry (2009). Sampling entails that some relevant variables will by chance have  $t^2 < c_\alpha^2$  in the given sample, so will not be selected, and hence conditional estimates will be biased away from the

origin because variables are retained only when  $t^2 \geq c_\alpha^2$ , often referred to as pre-test bias: see e.g., Judge and Bock (1978). Conversely, some irrelevant variables will have  $t^2 \geq c_\alpha^2$  (adventitiously significant) with probability less than or equal to  $\alpha(N - n)$  for  $n$  relevant variables. However, Hendry and Krolzig (2005) show that most of the selection bias can be corrected for relevant retained variables, at the cost of a small increase in their average conditional mean-square errors (MSEs). Such correction exacerbates the downward bias in the unconditional estimates of the relevant coefficients, so 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 for both their unconditional and conditional MSEs: see Castle et al. (2011). 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 due to testing irrelevant variables, but suffering some loss from not retaining relevant variables at large values of  $c_\alpha$ . 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  $\alpha$ , an issue examined in §3.1. Our desire to avoid underspecified GUMs derives from the converse costs of omitting relevant variables or breaks from a model, which can be large compared to those of selection: see e.g. Leeb and Pötscher (2005).

Our approach to detecting and removing multiple shifts includes an impulse indicator for every observation in the set of candidate regressors, called impulse-indicator saturation, as analyzed by Hendry et al. (2008a), and Johansen and Nielsen (2009). When the  $T$  impulse indicator variables are combined with the other candidate regressors, their total number must exceed the sample size. Johansen and Nielsen (2009) prove that under the null of no outliers or shifts, there is almost no loss of efficiency in testing for  $T$  impulse indicators for  $\alpha \leq 1/T$ , even in dynamic models. While surprising at first sight, retaining an impulse indicator when it is not needed merely removes one observation, which is all that happens on average. Thus, efficiency is of the order of  $100(1 - \alpha)\%$ . Monte Carlo experiments have confirmed the null distribution, and here we show that impulse-indicator saturation (denoted IIS) 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. Simulations compare our approach with stepwise regression, which can also handle  $N > T$ , and demonstrate the major gains from *Autometrics*.

The structure of the paper is as follows. Section 2 first illustrates IIS for the U.S. real interest rate, comparing to Bai and Perron (1998). Then §3 sketches the theory of impulse-indicator saturation under the null, and §4 reports a series of Monte Carlo experiments examining its ability to detect different forms and numbers of shifts in various DGPs, as well as comparing with stepwise regression. Section 5 revisits the real interest rate illustration for an extended sample period to show the potential benefits of IIS. Section 6 concludes.

## 2 Illustration of the U.S. *ex-post* real interest rate

Bai and Perron (2003) (hereafter BP) reconsider an application from Garcia and Perron (1996) who examine the evidence for structural change in the U.S. real interest rate (three-month treasury bill rate deflated by the CPI inflation rate). The data are quarterly from 1961:1 to 1986:3, and are recorded in Fig. 1. Garcia and Perron (1996) find evidence of 2 breaks (1973 and mid-1981). BP apply the Bai and Perron (1998) procedure with only a constant, and account for potential serial correlation using non-parametric adjustments. They find evidence of 3 breaks using the sequential procedure.<sup>1</sup> The break dates are 1966:4, 1972:3 and 1980:3, shown by vertical lines on Fig. 1.

---

<sup>1</sup>BP allow up to 5 breaks and use a trimming factor of 0.15 which corresponds to each segment having at least 15 observations. They find 2 breaks using BIC and modified BIC but conclude in favour of 3 breaks; see Table 1 in BP for details of the empirical results.

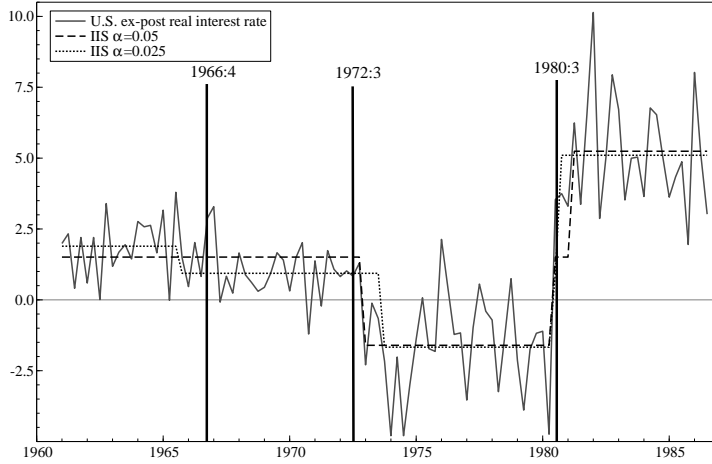


Figure 1: U.S. *ex-post* real interest rate, 1961:1–1986:3, with breaks determined by IIS at  $\alpha = 0.05$  (dashed line) and  $\alpha = 0.025$  (dotted line). Vertical lines are breaks found by BP.

We apply IIS to the real interest rate including just a constant as the regressor, using  $\alpha = 0.05, 0.025$ . IIS will retain individual impulse-indicators but these should occur in blocks corresponding to the segments between breaks. Hence, after applying IIS, we group the impulse-indicators with the same sign and similar magnitudes that occur sequentially (apart from ‘gaps’ in a segment) and form dummy variables taking the value 1 for the period covering the initial and last impulse-indicator in each group. At  $\alpha = 0.05$ , *Autometrics* retains 30 impulse-indicators, 15 of which are negative and occur in the period 1973:1–1980:2 (with a mean coefficient of  $-4.16$ ) and 15 are positive, occurring in the period 1981:2–1986:2 (with a mean coefficient of  $4.94$ ). Fig. 1 records the two breaks by a dashed line, and the results correspond to the findings of Garcia and Perron (1996). Applying IIS at  $\alpha = 0.025$  also results in 30 impulse-indicators being retained. The first 6 correspond to a segment dated 1962:2–1965:3, the second 6 to 1973:4–1980:2 and the third 18 to 1980:4–1986:2. These breaks are recorded by the dotted line in Fig. 1. The results are similar to BP, although the first two dates are a little earlier, but within the uncertainty bands that BP report. Hence, IIS is able to replicate the findings of those previous studies in determining evidence of structural change. Section 5 revisits this application for a longer sample period.

### 3 Impulse-indicator saturation

Impulse-indicator saturation adds an impulse indicator for every observation to the candidate regressor set, entered (in the simplest case) in blocks of  $T/2$ , with the significant outcomes retained. First, add half the impulse indicators, record the significant outcomes, then drop that set of impulse indicators and add the other half, recording significant outcomes again. These first two steps correspond to ‘dummying out’  $T/2$  observations for estimation, since impulse indicators are mutually orthogonal. Now combine the recorded impulse indicators and select those that remain significant. Then under the null of no outliers or breaks,  $\alpha T$  impulse 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., unity). The theory generalizes to more, and unequal, splits, as well as dynamic models, see Johansen and Nielsen (2009). The theory of IIS is under the null of no outliers, but with the aim of detecting and removing any outliers, data contamination, and location shifts.

*Autometrics* uses its general expansion and contraction algorithm even though impulse indicators are orthogonal, and generally tries several block divisions. Its null-distribution theory has not been

developed, so to assess its operational characteristics under the null, we first conduct a group of Monte Carlo simulations. The experimental design builds on Castle et al. (2009) with  $T = 100$ :

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

$$x_{i,t} = \rho x_{i,t-1} + v_{i,t} \quad v_{i,t} \sim \text{IN} [0, 1], \quad i = 1, \dots, 10, \quad (2)$$

$$\epsilon_t \sim \text{IN} [0, 1], \quad t = 1, \dots, T, \quad (3)$$

where  $\mathbf{x}'_t = (x_{1,t}, \dots, x_{10,t})$  are fixed across replications. Equations (1)–(3) specify 5 different DGPs, indexed by  $n = 1, \dots, 5$ , each having  $n + 2$  relevant variables, with  $\beta_1 = \dots = \beta_n \neq 0$ , plus the intercept and lagged dependent variable (LDV), and  $10 - n$  irrelevant variables ( $\beta_{n+1} = \dots = \beta_{10} = 0$ ). Throughout we set  $\beta_0 = 1$ ,  $\rho = 0.5$  and  $\gamma = 0.5$ . For the  $n = 1, \dots, 5$  experiments,  $\beta_1 = \dots = \beta_n = \psi/\sqrt{T}$ , where  $\psi = 2, \dots, 6$ , such that all relevant exogenous variables in each experiment will have the same population t-value.

The GUM is the same for all 5 DGPs:  $\{1, y_{t-1}, x_{1,t} \dots x_{10,t}\}$ .  $M = 1000$  replications are undertaken throughout, unless otherwise noted. Gauge and potency denote the empirical null retention frequency and average non-null retention frequency. We calculate the potency over the exogenous regressors only, not counting the intercept and LDV as these are highly significant (so  $N = 10$ ):

$$\begin{aligned} \text{retention rate } \tilde{p}_k &= \frac{1}{M} \sum_{i=1}^M 1_{(\tilde{\beta}_{k,i} \neq 0)}, \quad k = 1, \dots, N; \\ \text{potency} &= \frac{1}{n} \sum_{k=1}^n \tilde{p}_k, \\ \text{gauge} &= \frac{1}{N-n} \left( \sum_{k=n+1}^N \tilde{p}_k \right) \end{aligned}$$

where  $\tilde{\beta}_{k,i}$  denotes the OLS coefficient on  $x_{k,t}$  in replication  $i$  if selected (0 otherwise), and  $1_{(\tilde{\beta}_{k,i} \neq 0)}$  is the indicator variable, equal to unity when the argument is true and zero otherwise.

$\alpha$	1%		0.1%	
	no IIS	IIS	no IIS	IIS
ave. gauge (across all $\psi$ )	1.22	1.65	0.20	0.10
ave. potency $\psi = 2$	34.33	31.59	12.69	9.70
$\psi = 3$	78.14	73.03	50.54	42.44
$\psi = 4$	97.33	94.92	87.10	83.49
$\psi = 5$	99.78	99.44	98.62	98.27
$\psi = 6$	100.0	99.98	99.88	99.83

Table 1: Gauge and potency for exogenous regressors ( $\times 100$ ) for selection averaged across all  $n = 1, \dots, 5$  experiments (and  $\psi = 2, \dots, 6$  for gauge) with and without IIS.

Table 1 records the gauge averaged across all  $n = 1, \dots, 5$  and  $\psi = 2, \dots, 6$  experiments, and potency averaged across all  $n = 1, \dots, 5$  experiments (no diagnostic testing is undertaken in these *Autometrics* experiments). The gauge is too large at  $\alpha = 1\%$  with IIS, as 1 impulse indicator is retained on average, inducing a slight downward bias in the residual standard deviation, but is equal to  $\alpha$  at 0.1% with IIS, although too large without. The average potencies are similar to single t-test powers,  $\Pr(t_{\psi=0} \geq c_\alpha | \psi)$ , even with IIS, although IIS lowers potency, particularly at tight  $\alpha$ .

We also calculate conditional and unconditional MSEs as:

$$\begin{aligned} \text{UMSE}_k &= \frac{1}{M} \sum_{i=1}^M \left( \tilde{\beta}_{k,i} - \beta_k \right)^2, \\ \text{CMSE}_k &= \frac{\sum_{i=1}^M \left[ \left( \tilde{\beta}_{k,i} - \beta_k \right)^2 \cdot 1_{(\tilde{\beta}_{k,i} \neq 0)} \right]}{\sum_{i=1}^M 1_{(\tilde{\beta}_{k,i} \neq 0)}}, \quad (\beta_k^2 \text{ when } \sum_{i=1}^M 1_{(\tilde{\beta}_{k,i} \neq 0)} = 0). \end{aligned} \quad (4)$$

In (4), the unconditional MSE (UMSE) substitutes zeros when a variable is not selected, and the conditional MSE (CMSE) is computed over retained variables only.

Figure 2 records the ratio of bias-corrected MSEs without IIS to that with IIS. At  $\alpha = 1\%$ , where 1 impulse indicator is retained on average, the ratios are below unity: the MSEs are larger with than without IIS, particularly for irrelevant variables. At tight significance levels ( $\alpha = 0.1\%$ ), so few dummies are retained that it has no systematic impact on the MSE. Bias correction substantially reduces the MSEs for irrelevant variables at a small cost of increasing the MSEs for relevant variables.

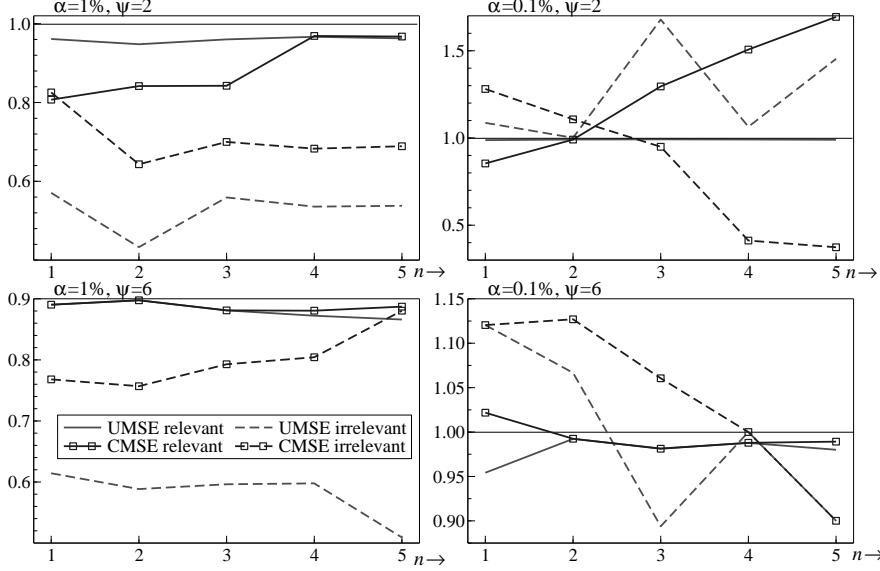


Figure 2: Ratios of bias-corrected MSEs without IIS to with, averaged across all relevant (solid lines) and irrelevant (dashed lines) variables. Top panels correspond to  $\psi = 2$  and bottom panels  $\psi = 6$ .

### 3.1 Impulse-indicator saturation in a fat-tailed distribution

As impulse-indicator saturation is designed to detect outliers and location shifts, we assess its impact for a fat-tailed error distribution, using the Student-t distribution with 3 degrees of freedom. The experiment is identical to (1) and (2), but (3) becomes  $\epsilon_t \sim t_3$ .

Table 2 records both average gauges across all  $n = 1, \dots, 5$  and  $\psi = 2, \dots, 6$  experiments, and potency averaged over the  $n = 1, \dots, 5$  experiments for  $t_3$ , with diagnostic testing, without, and with IIS at  $\alpha = 1\%$  (approximately  $1/T$  here) and the tighter  $\alpha = 0.1\%$ . *Autometrics* checks normality in its batch of mis-specification tests. If it rejects, the p-value of later normality tests is reduced, but the program tries to return to the original p-value at a later stage in selection, and may retain irrelevant variables which help ensure the diagnostic test is passed. We also calculate the average retention probability of the irrelevant variables not counting impulse indicators.<sup>2</sup>

If diagnostic testing is applied without IIS when 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 (4% for  $\alpha = 0.1\%$ ), as *Autometrics* retains additional irrelevant variables to try to improve normality. Omitting diagnostic testing substantially improves the gauge when IIS is off, but lowers potency. When IIS is on, the impulse indicators capture much of the non-normality, reducing the gauge for other regressors

<sup>2</sup>The definition of gauge is ambiguous when retained impulse indicators could be 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 correspond to extreme observations, so should be retained as contributing to potency.

IIS		no	no	yes	yes	no	no	yes	yes
diagnostics		yes	no	yes	no	yes	no	yes	no
$\alpha$		1%				0.1%			
gauge		5.66	1.65	5.79	5.77	3.92	0.34	1.63	1.53
gauge*		—	—	2.11	2.08	—	—	0.49	0.23
potency	$\psi = 2$	16.87	11.80	20.63	20.57	10.16	3.46	4.54	3.80
	$\psi = 3$	34.90	31.37	49.65	49.59	17.87	11.89	14.66	13.94
	$\psi = 4$	58.85	57.73	79.21	79.20	34.81	31.33	39.53	38.78
	$\psi = 5$	78.99	79.05	94.07	94.10	57.27	56.44	67.77	67.57
	$\psi = 6$	90.68	90.74	98.67	98.67	76.78	76.60	88.49	88.39

Table 2: Gauge and potency in % for  $t_3$  error distribution, averaged across all  $\psi = 2, \dots, 6$  and  $n = 1, \dots, 5$  experiments for gauge, and  $n = 1, \dots, 5$  experiments for potency; gauge\* denotes the gauge not counting indicators.

substantially. While the critical values used in the selection algorithm would be incorrect for a  $t_3$ -distribution, the ‘near normality’ of the resulting distribution after IIS leads to the gauge being closer to, although still larger than,  $\alpha$ .

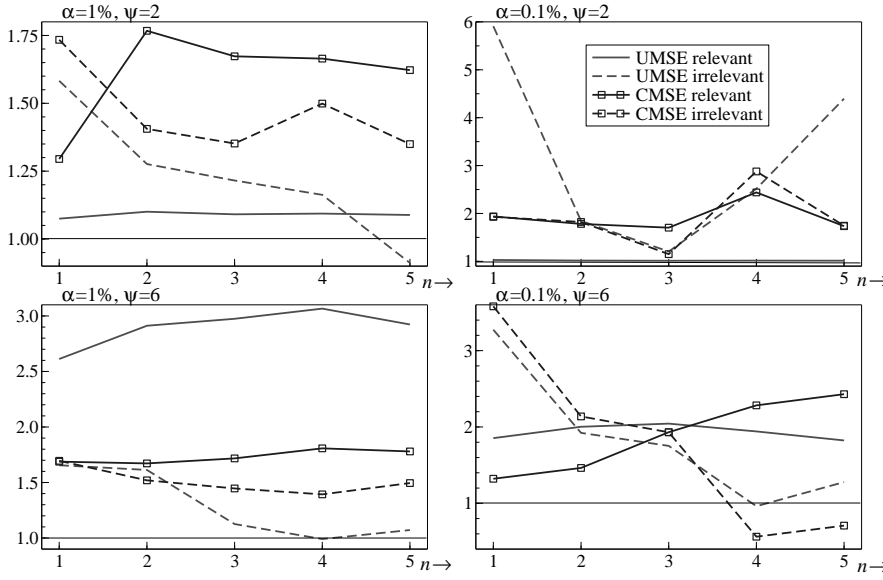


Figure 3: Ratios of MSEs without IIS to with for a  $t_3$ -distribution with bias correction and no diagnostic testing, averaged across all relevant (solid lines) and irrelevant (dotted lines) variables. Top panels correspond to  $\psi = 2$  and bottom panels  $\psi = 6$ .

In almost all cases, potency is higher when IIS is applied to account for the fat tails: the loss in potency at  $\alpha = 0.1\%$  for  $\psi = 2, 3$  relative to no IIS with diagnostic tracking is due to the latter’s large gauge: potencies are not gauge corrected. At  $\alpha = 1/T = 1\%$ , *Autometrics* performs well in its standard operational mode of IIS and diagnostic testing.

Figure 3 records the ratio of MSEs without IIS to with IIS for  $t_3$  errors. Most ratios are larger than unity, demonstrating the benefits of IIS by smaller MSEs for the coefficient estimates of the retained regressors at both significance levels (the MSE ratios that are not bias corrected are larger). Applying IIS leads to the coefficient estimates for the retained variables being closer to their DGP values.

To evaluate the impact of IIS, Fig. 4 compares the conditional distributions for  $\tilde{\beta}_0, \dots, \tilde{\beta}_{10}$  when  $n = 5$ , corresponding to  $\beta_1, \dots, \beta_5 = 1$  and  $\beta_6, \dots, \beta_{10} = 0$  (now with  $\beta_0 = 5$  and  $\rho = 0$ ), with the IIS distributions superimposed. For the relevant variables, the distributions are similar, although the

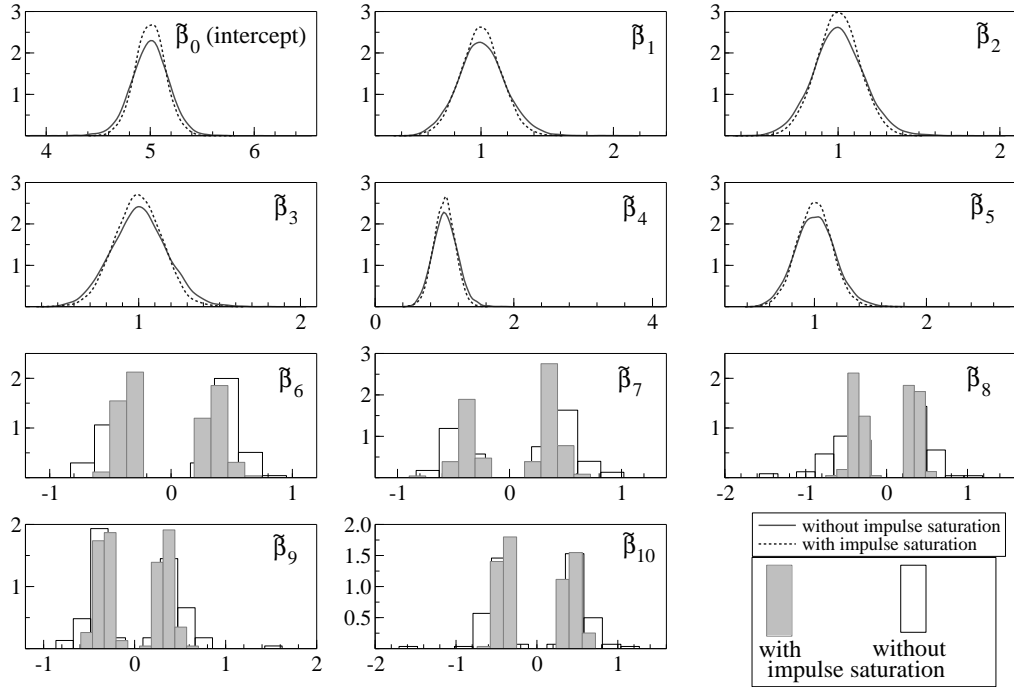


Figure 4: Distributions of estimates conditional on retention with (dark) and without (light) IIS for the  $n = 5$  experiment with a  $t_3$  error distribution at  $\alpha = 1\%$ ,  $M = 10000$ .

distributions without IIS have slightly fatter tails. The long tails for the non-IIS conditional distributions of the coefficient estimates for the irrelevant variables relative to the IIS outcomes are evident. Thus, if irrelevant variables are retained in the presence of fat tails, IIS implies that the reported coefficient estimates are smaller than otherwise, which bias correction further downweights, providing insurance against falsely substantive irrelevant variables.

## 4 Selecting regressions when there are multiple breaks

Having established that including  $T$  impulse indicators need not induce a large efficiency loss when they are irrelevant, but help ‘correct’ in fat-tailed distributions, we now apply IIS to detect outliers and locations shifts when such features do in fact occur. The generality of IIS allows it to detect many shifts, and in simulation experiments, we consider up to 20 in 100 observations, only one, and several intermediate settings. Moreover, the shifts can be at the start or end of the sample, as there is no need to reserve a percentage as with (say) Bai and Perron (1998). Finally, regressors, lags, and functional forms can be selected jointly with IIS, although the last two are not the focus here. All experiments have  $T = 100$  using  $M = 1000$  replications.

Hendry and Santos (2010) present analytic ‘power’ calculations for IIS for a single outlier and a single level shift in a marginal process, assuming normality, but extending such results to multiple breaks is problematic as the detectability of any given break depends on that of all others, and at best could provide conditional probabilities: the significance of each indicator depends inversely on the estimated residual standard deviation, which is upward biased if any shifts are not detected. The downward bias under the null due to selection is easily corrected: see Johansen and Nielsen (2009).



#### 4.1 IIS for breaks in the mean of a location-scale model

The first set of experiments examines the detectability of various forms of location shift in the simplest setting, but for a range of magnitudes, forms and timings of breaks. The DGPs are listed in Table 3, where  $I_t$  is an impulse indicator for observation  $t$ .

DGP:Bc	$y_t = \delta + \gamma (I_{81} + \dots + I_{100}) + u_t,$
DGP:B20	$y_t = \delta + \gamma (I_1 + I_6 + I_{11} + \dots + I_{96}) + u_t,$
DGP:MBc	$y_t = \delta + \gamma (I_1 + I_2 + I_3 + I_4 + I_{24} + \dots + I_{27} + I_{49} + \dots + I_{52} + I_{74} + \dots + I_{77} + I_{97} + \dots + I_{100}) + u_t,$
DGP:Bct	$y_t = \delta + \gamma (I_{81} + \dots + I_{100}) + 0.02t + u_t.$
DGP:BL	$y_t = \gamma (I_{81} + \dots + I_{100}) + 0.5y_{t-1} + u_t, \quad y_0 = 0,$
DGP:BLc	$y_t = 2 + \gamma (I_{81} + \dots + I_{100}) + 0.5y_{t-1} + u_t, \quad y_0 = 0,$
	$u_t \sim \text{IN}[0, 1]; \quad \delta = 0, 1 \text{ as noted below}$
GUM:Ic	$y_t$ on 1 (free) and $T$ indicators: DGP:Bc, B20, MBc;
GUM:Ict	$y_t$ on 1 (free), $T$ indicators, and trend: DGP:Bct
GUM:IcL	$y_t$ on 1 (forced and free), $T$ indicators, and $y_{t-1}$ : DGP:BL, BLc.

Table 3: DGPs and GUMs for location-scale and autoregressions with breaks.

Thus, DGP:Bc has a single break in the mean starting at  $T = 81$ , whereas DGP:B20 has 20 breaks in the mean, starting at  $T = 1$ , which are equally spread for simulation convenience, but that knowledge is not used, and DGP:MBc has 5 breaks, each of length four, again equally spread. Next, DGP:Bct adds a trend to DGP:Bc. Third, DGP:BL has a break in the mean, starting at  $T = 81$ , in a stationary autoregression with a zero mean before, and finally DGP:BLc has the same break in the mean as DGP:Bc but in a stationary autoregression. *Autometrics* was started from the GUMs listed in Table 3, where ‘forced’ entails that the relevant variable cannot be eliminated by selection, and 1 denotes a constant.

	$\gamma = 0$	$\gamma = 1$	$\gamma = 2$	$\gamma = 3$	$\gamma = 4$	$\gamma = 5$
	<i>Autometrics</i> , constant free, $\alpha = 1\%$					
	DGP:Bc					
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					
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 4: *Autometrics* for IIS in location-scale DGPs with breaks at the end and multiple breaks. Constant free,  $\delta = 0$ .

Table 4 first reports in detail the results for DGP:Bc and DGP:B20 when  $\delta = 0$  for  $\gamma = 1$  up to  $\gamma = 5$ . It is much easier to detect a single break of length 20 than twenty breaks of 1 period when  $\gamma$  is small, but the potencies rapidly rise towards unity in both cases as  $\gamma$  grows to 5. While 20 ‘shifts’ in a sample of 100 is unlikely in practice, the ability to find them is encouraging, as data contamination on that scale is quite possible. The gauge falls for intermediate values of  $\gamma$  as ‘missed’ breaks augment the error variance, and reduce the probability of retaining impulses for observations without a break. This effect vanishes once all breaks are detected (here by  $\gamma \geq 5$ ). Five breaks of length 4, as in DGP:MBc, is realistic, and has a potency that lies between the two cases reported in Table 4.

## 4.2 Comparisons with stepwise regression

Table 5 compares *Autometrics* with stepwise regression to highlight the improvements due to path search, now with  $\delta = 1$  (DGP:Bc is omitted because it is almost the same as DGP:MBc). The potency of *Autometrics* rises rapidly with  $\gamma$  in all four DGPs, with almost all breaks detected by  $\gamma = 5$ . *Autometrics* highly outperforms stepwise regression, a pattern that occurred in every experiment, so we do not report the latter henceforth (stepwise results are available on request).

	$\gamma = 3$	$\gamma = 4$	$\gamma = 5$	$\gamma = 3$	$\gamma = 4$	$\gamma = 5$
	<i>Autometrics</i> , constant forced, $\alpha = 1\%$					
	DGP:Bct			DGP:MBc		
gauge %	4.7	1.6	1.1	0.4	0.7	1.0
potency %	29.4	68.9	92.2	38.8	78.4	96.5
	Stepwise regression, constant forced, $\alpha = 1\%$					
	DGP:Bct			DGP:MBc		
gauge %	0.7	0.4	0.2	0.1	0.0	0.1
potency %	6.2	6.2	5.9	14.2	16.9	18.2

Table 5: *Autometrics* and stepwise model selection for IIS in location-scale and trend DGPs with breaks at the end and multiple breaks, with forced constant and  $\delta = 1$ .

## 4.3 IIS for breaks in the mean of a stationary autoregression

Table 6 reports the simulation results for breaks in the mean of a stationary autoregression.<sup>3</sup>

	$\gamma = 5$	$\gamma = 8$	$\gamma = 10$	$\gamma = 5$	$\gamma = 8$	$\gamma = 10$
	<i>Autometrics</i> , 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
	<i>Autometrics</i> , constant forced, $\alpha = 1\%$					
	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 6: *Autometrics* for IIS in stationary autoregressive DGPs with breaks at end, comparing free and forced constants.

The treatment of the constant matters greatly for *Autometrics*. When the constant is free in DGP:BLc it is often not retained, so the selected model has a unit root with one or two indicators at the start of the break. While a good approximation in terms of fit, that is a poor choice for the LDGP. With a forced constant, however, *Autometrics* is able to detect the correct model in both cases, with the intercept close to zero in the former.

## 4.4 IIS 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. We consider the experimental designs listed in Table 7. In all cases, the intercept is forced to be retained. DGUM has  $\Delta y_t$  as the dependent variable, while the regressors include  $y_{t-1}$ . After

<sup>3</sup>Because discarding initial observations is simpler to code than starting from the unconditional mean, DGP:BLc is started from  $y_{-45} = 0$ , then the first 45 observations are discarded, so the actual data are virtually identical to starting from  $y_0 = 4$ .

DGP:IUc	$y_t = 0.2 + \gamma I_{81} + y_{t-1} + u_t,$
DGP:BUc	$y_t = 0.2 + \gamma (I_{81} + \dots + I_{100}) + y_{t-1} + u_t,$
DGP:MIUc	$y_t = 0.2 + \gamma (I_1 + I_{24} + I_{49} + I_{74} + I_{97}) + y_{t-1} + u_t,$
DGP:MBUc	$y_t = 0.2 + \gamma (I_1 + \dots + I_4 + I_{24} + \dots + I_{27} + I_{49} + \dots + I_{52} + I_{74} + \dots + I_{77} + I_{97} + \dots + I_{100}) + y_{t-1} + u_t,$
	$u_t \sim \text{IN}[0, 1], y_{-100} = 0, t = -99, \dots, 0, 1, \dots, 100$
GUM:ILct	$y_t$ on 1 and $T$ indicators, $y_{t-1}$ , and trend, $t = 1, \dots, 100$ ;
DGUM:ILct	$\Delta y_t$ on 1, $T$ indicators, $y_{t-1}$ , and trend, $t = 1, \dots, 100$ .

Table 7: DGPs and GUMs for unit-root experiments.

	$\gamma = 4$	$\gamma = 5$	$\gamma = 4$	$\gamma = 5$	$\gamma = 4$	$\gamma = 5$	$\gamma = 4$	$\gamma = 5$
	DGP:IUc		DGP:BUc		DGP:MIUc		DGP:MBUc	
	<i>Autometrics</i> , $\alpha = 1\%$ , constant forced, starting from GUM:ILct							
gauge %	1.3	1.3	0.9	0.9	1.1	1.1	1.9	1.6
potency %	95.7	99.1	37.8	52.5	79.3	93.3	57.2	64.8
	<i>Autometrics</i> , $\alpha = 1\%$ , constant forced, starting from DGUM:ILct							
gauge %	1.6	1.6	1.0	1.3	1.5	1.5	0.9	1.0
potency %	96.6	99.4	59.7	90.4	89.7	98.2	71.4	94.7

Table 8: *Autometrics* results for IIS in a unit-root DGP with breaks.

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 8 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 in the level formulation.

#### 4.5 Autoregressions with regressors, without and with breaks

In these experiments, the extra regressors in the DGP are specified as:

$$\mathbf{x}'_t \boldsymbol{\beta} = \sum_{i=1}^4 \beta^* (x_{i,t} - x_{i,t-1}), \quad (5)$$

for  $\beta^* = (T[1 - \rho])^{-1/2} \beta$  (scaled for sample size and autoregression). We vary  $\beta$  across experiments:  $\beta = \{2.4, 3.2, 4.0\}$ .

The regressors are generated with unit variance and  $\rho$  fixed at 0.9 with  $x_{i,0} = 0$  in:

$$x_{i,t} = \rho x_{i,t-1} + v_{i,t}, \quad i = 1, \dots, 10, \quad v_{i,t} \sim \text{IN}[0, (1 - \rho)^2]. \quad (6)$$

Table 9 lists the designs. All 10 regressors and their lags are added to the GUM, so there are 8 relevant regressors (four at lag zero and four at lag one), and 12 irrelevant. When  $\gamma = 0$ , the experiments are under the null of no break and we can omit the ‘B’ from the DGP label.

This setup is very challenging because the regressors are highly correlated. Moreover, they enter the DGP in differences, but are formulated in levels in the GUM. The worst gauges are around 10% for the unit root DGP with a trend in the initial model and no breaks, see Table 10. When  $\rho$  in (6) is reduced from 0.9 to 0.1, these gauges reduce to about 5%. Then formulating the GUM in differences reduces it further.

DGP:BLcx	$y_t = 2 + \gamma(I_{81} + \dots + I_{100}) + 0.5y_{t-1} + \mathbf{x}_t'\boldsymbol{\beta} + u_t, \quad y_0 = 0, \quad t = 1, \dots, 100,$
DGP:BUcx	$y_t = 0.2 + \gamma(I_{81} + \dots + I_{100}) + y_{t-1} + \mathbf{x}_t'\boldsymbol{\beta} + u_t, \quad y_{-100} = 0, \quad t = -99, \dots, 0, \dots, 100,$
GUM:Lcx	$y_t$ on 1 (forced), $y_{t-1}, x_{1,t} \dots x_{10,t}, x_{1,t-1} \dots x_{10,t-1}, t = 1, \dots, 100.$
GUM:Lcxt	GUM:Lcx with trend.
GUM:ILcx	GUM:Lcx with $T$ indicators.
GUM:ILcxt	GUM:Lcx with $T$ indicators and trend.

Table 9: DGPs and GUMs for autoregressive experiments with regressors

	$\gamma = 0$			$\gamma = 5$			$\gamma = 10$	
	<i>Autometrics</i> , Constant forced, $\alpha = 1\%$							
	$\beta = 2.4$	3.2	4.0	$\beta = 2.4$	3.2	4.0	$\beta = 2.4$	4.0
	DGP:Lcx, GUM:Lcx			DGP:BLcx, GUM:ILcx				
gauge %	4.3	3.2	2.4	2.2	2.1	2.2	1.9	2.0
potency %	58.3	86.3	97.8	35.0	38.0	43.2	64.3	76.6
	DGP:Ucx, GUM:Lcx			DGP:BUcx, GUM:ILcx				
gauge %	3.9	3.3	2.8	2.5	2.8	2.4	2.5	1.5
potency %	59.5	85.3	96.8	33.1	39.2	45.0	60.7	67.5
	DGP:Ucx, GUM:Lcxt			DGP:BUcx, GUM:ILcxt				
gauge %	11.9	10.8	9.2	3.0	3.1	3.0	3.5	3.3
potency %	55.8	79.3	92.3	36.6	42.6	47.1	64.3	69.1

Table 10: *Autometrics* results for stationary and unit-root autoregressive DGPs. Potency and gauge are over regressors for  $\gamma = 0$ ; over regressors and impulses otherwise.

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 as seen in Table 10 for  $\gamma = 5$  and  $\gamma = 10$ .

## 5 Revisiting the U.S. *ex-post* real interest rate

We extend the sample to 1947:2–2009:3, and compare BP and IIS over 250 observations. The sequential BP procedure is applied allowing for up to 8 breaks, with a trimming factor of 0.10 which corresponds to each segment having at least 25 observations, and allows for serial correlation in the errors and different variances of the residuals across segments. IIS identifies location shifts as blocks of impulse indicators with the same signs and similar magnitudes, and retains individual impulse indicators for large outliers. Both BP and IIS are applied at  $\alpha = 0.01, 0.025$  and  $0.05$ . IIS is also applied at  $\alpha = 0.001$ , when 0.25 irrelevant impulse indicators will be retained on average under the null.

The results for BP are reported in Tables 11; IIS is in Table 12. As IIS does not impose a minimum segment length, some of the segments are capturing outliers as opposed to mean shifts. If consecutive impulse indicators differ significantly in magnitude, then individual impulse indicators are retained rather than imposing a step shift, particularly so at the beginning of the sample.

The breaks are shown in Fig. 5. The BP procedure, shown in the top panel, detects very few breaks, not finding most of those reported on the shorter sample (1966:4, 1972:3 and 1980:3 were found in the

BP at 5%	80:2 [80:1-83:2], 86:3 [85:2-87:3], 01:4 [95:4-03:2]
BP at 2.5%	80:2 [80:1-83:2], 86:3 [84:3-87:1]
BP at 1%	same as breaks at 2.5%

Table 11: Breaks for U.S. ex-post real interest rate detected using the sequential procedure of BP; 1947:2–2009:3; 95% confidence intervals for break dates in brackets. BIC selects 5 breaks and modified BIC selects 2 breaks.

IIS 5%	75 impulse indicators retained: 7 segments, 15 outliers
IIS 2.5%	33 impulse indicators retained: 4 segments, 12 outliers
segments	74:1-80:2, 81:2-86:1, 03:1-03:3, 07:4-09:3
outliers	47:3, 47:4, 48:2, 50:2, 50:3, 50:4, 51:1, 51:4, 58:1, 89:3, 05:3, 08:4
IIS 1%	21 impulse indicators retained: 2 segments, 11 outliers
segments	74:1-80:2, 81:4-86:1
outliers	47:3, 47:4, 48:2, 50:3, 50:4, 51:1, 51:4, 58:1, 05:3, 08:2, 08:4
IIS 0.1%	49 impulse indicators retained: 5 segments, 10 outliers
segments	73:1-80:3, 81:2-86:2, 89:1-89:3, 03:1-05:2, 07:4-09:3
outliers	47:3, 47:4, 48:2, 50:3, 50:4, 51:1, 51:4, 98:1, 05:3, 08:4

Table 12: IIS results for U.S. ex-post real interest rate, 1947:2–2009:3

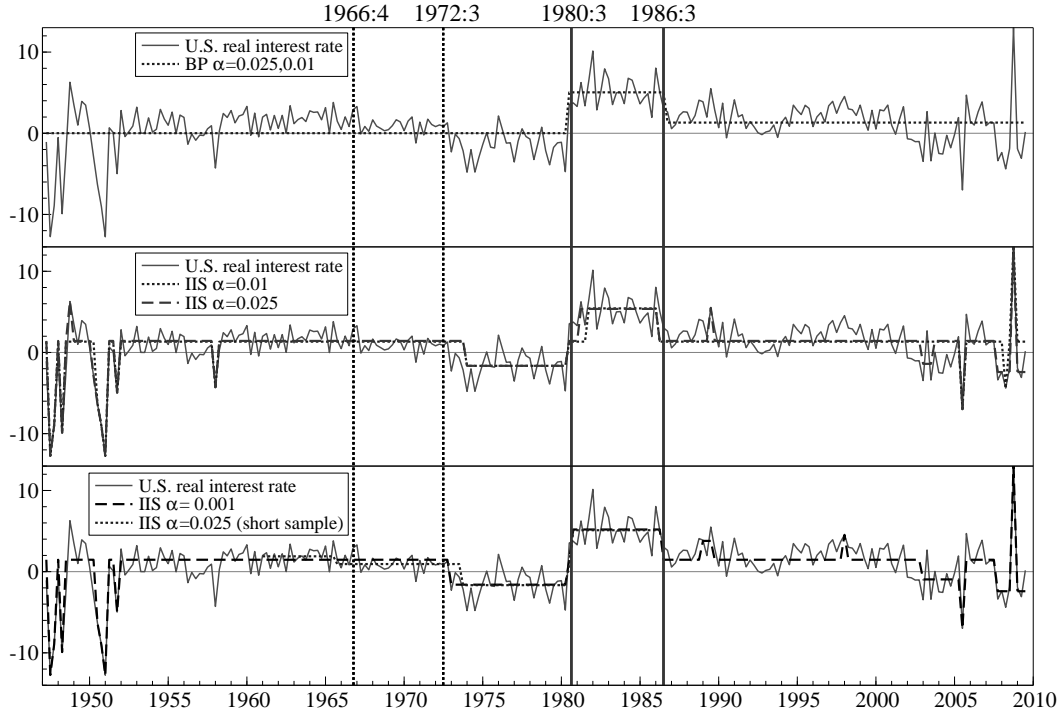


Figure 5: U.S. real interest rate, 1947:2–2009:3, with breaks determined by BP (top panel) and IIS (middle and bottom panels). BP breaks from extended sample are shown as solid vertical lines; solid dotted lines are from short sample.

shorter sample, and are marked by vertical lines, as is 1986:3 which was found with the extended data).

IIS is reported in the middle panel Fig. 5. IIS detects both the large outliers and breaks at the beginning and end of the sample, where BP requires a reserved sample percentage. At the tight significance level  $\alpha = 0.001$ , reported in the bottom panel of Fig. 5, where the BP breaks are shown as vertical lines, IIS detects the location shifts despite the outliers, while controlling the null retention frequency at 0.25 of an indicator on average. In fact, more impulse indicators are retained than at the 1% significance level

— a consequence of the expanding and contracting searches.

This application demonstrates the ability of IIS to detect both location shifts and outliers, where the latter are close to the start and end of the sample, and need removed to ‘reveal’ the step shifts. The further advantage of *Autometrics*, namely that variables can be jointly selected while detecting outliers and breaks, was illustrated above. The final stage of the procedure — replacing individual impulse indicators with step dummies — is currently not automated, but an algorithm could do so.

## 6 Conclusion

*Autometrics* with impulse-indicator saturation can select variables jointly with tackling multiple breaks at unknown times, and can handle fat-tailed distributions. It performs well in detecting breaks in location-scale, location-scale-trend, stationary autoregressions (with one exception), and in a unit-root DGP. In comparison to a single break, multiple breaks are harder to detect in a DGP without a trend, but easier in a DGP with a trend. However, *Autometrics* performs poorly for a break in a stationary autoregression with an intercept that is not forced, easily corrected by forcing it. In the simulations, *Autometrics* controls its null retention frequency quite close to the nominal significance level  $\alpha$ , and its retention of relevant variables is near the corresponding test power at tight  $\alpha$ .

The empirical illustration to U.S. real interest rates compared IIS with the procedure in Bai and Perron (1998), and found a similar number and timing of breaks over their sample period. However, extending the sample to 1947:2–2009:3, covering most of the post-war period, revealed substantial benefits to IIS as there are breaks or outliers near the start and end of the sample as well as other shifts.

Overall, *Autometrics* IIS, allows multiple breaks to be detected, using both expanding and contracting searches because there are more candidate regressors than observations, without losing much efficiency when there are no breaks.

## Acknowledgements

Financial support from the British Academy is gratefully acknowledged by the first author. This research was supported in part by grants from the Open Society Institute and the Oxford Martin School. We are indebted to the Editors and two anonymous referees of the *Journal of Econometrics* for helpful comments on previous versions.

## References

- Bai, J., Perron, P., 1998. Estimating and testing linear models with multiple structural changes. *Econometrica* 66, 47–78.
- Bai, J., Perron, P., 2003. Computation and analysis of multiple structural change models. *Journal of Applied Econometrics* 18, 1–22.
- Castle, J. L., Doornik, J. A., Hendry, D. F., 2011. Evaluating automatic model selection. *Journal of Time Series Econometrics* 3 (1), DOI: 10.2202/1941–1928.1097.
- Castle, J. L., Qin, X., 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., Shephard, N. (Eds.), 2009. *The Methodology and Practice of Econometrics*. Oxford University Press, Oxford.
- Doornik, J. A., 2008. Encompassing and automatic model selection. *Oxford Bulletin of Economics and Statistics* 70, 915–925.
- Doornik, J. A., 2009a. Autometrics. In: Castle and Shephard (2009), pp. 88–121.
- Doornik, J. A., 2009b. *Object-Oriented Matrix Programming using Ox*, 7th Edition. Timberlake Consultants Press, London.
- Garcia, R., Perron, P., 1996. An analysis of the real interest rate under regime shifts. *Review of Economics and Statistics* 78, 111–125.
- Hendry, D. F., 2009. The methodology of empirical econometric modeling: Applied econometrics through the looking-glass. In: Mills, T. C., Patterson, K. D. (Eds.), *Palgrave Handbook of Econometrics*. Palgrave MacMillan, Basingstoke, pp. 3–67.
- Hendry, D. F., Doornik, J. A., 2009. *Empirical Econometric Modelling using PcGive: Volume I*. Timberlake Consultants Press, London.
- Hendry, D. F., Johansen, S., 2010. Model selection when forcing retention of theory variables. Unpublished paper, Economics Department, University of Oxford.
- Hendry, D. F., Johansen, S., Santos, C., 2008a. Automatic selection of indicators in a fully saturated regression. *Computational Statistics* 33, 317–335, erratum, 337–339.
- Hendry, D. F., Krolzig, H.-M., 2005. The properties of automatic Gets modelling. *Economic Journal* 115, C32–C61.
- Hendry, D. F., Marcellino, M., Mizon, G. E. (Eds.), 2008b. Encompassing. *Oxford Bulletin of Economics and Statistics*, Special Issue.
- Hendry, D. F., Nielsen, B., 2007. *Econometric Modeling: A Likelihood Approach*. Princeton University Press, Princeton.
- Hendry, D. F., Santos, C., 2010. An automatic test of super exogeneity. In: Watson, M. W., Bollerslev, T., Russell, J. (Eds.), *Volatility and Time Series Econometrics*. Oxford University Press, Oxford, pp. 164–193.
- Hoover, K. D., Perez, S. J., 1999. Data mining reconsidered: Encompassing and the general-to-specific approach to specification search. *Econometrics Journal* 2, 167–191.
- Johansen, S., Nielsen, B., 2009. An analysis of the indicator saturation estimator as a robust regression estimator. In: Castle and Shephard (2009), pp. 1–36.
- Judge, G. G., Bock, M. E., 1978. *The Statistical Implications of Pre-Test and Stein-Rule Estimators in Econometrics*. North Holland Publishing Company, Amsterdam.
- Leamer, E. E., 1983. Let's take the con out of econometrics. *American Economic Review* 73, 31–43.
- Leeb, H., Pötscher, B. M., 2005. Model selection and inference: Facts and fiction. *Econometric Theory* 21, 21–59.
- Schwarz, G., 1978. Estimating the dimension of a model. *Annals of Statistics* 6, 461–464.