Castle, Jennifer; Doornik, Jurgen A.; Hendry, David F.

Article

Selecting a model for forecasting

Econometrics

Provided in Cooperation with:
MDPI – Multidisciplinary Digital Publishing Institute, Basel

Suggested Citation: Castle, Jennifer; Doornik, Jurgen A.; Hendry, David F. (2021) : Selecting a model for forecasting, Econometrics, ISSN 2225-1146, MDPI, Basel, Vol. 9, Iss. 3, pp. 1-35, https://doi.org/10.3390/econometrics9030026

econometrics

MDPI

Article

Selecting a Model for Forecasting

Jennifer L. Castle [1,*], Jurgen A. Doornik [2] and David F. Hendry [2]

[1] Magdalen College and Climate Econometrics, University of Oxford, High Street, Oxford OX1 4AU, UK
[2] Nuffield College, Climate Econometrics and Institute for New Economic Thinking at the Oxford Martin School, University of Oxford, Nuffield College, New Road, Oxford OX1 1NF, UK; jurgen.doornik@nuffield.ox.ac.uk (J.A.D.); david.hendry@nuffield.ox.ac.uk (D.F.H.)
* Correspondence: jennifer.castle@magd.ox.ac.uk; Tel.: +44-01865-276067

Abstract: We investigate forecasting in models that condition on variables for which future values are unknown. We consider the role of the significance level because it guides the binary decisions whether to include or exclude variables. The analysis is extended by allowing for a structural break, either in the first forecast period or just before. Theoretical results are derived for a three-variable static model, but generalized to include dynamics and many more variables in the simulation experiment. The results show that the trade-off for selecting variables in forecasting models in a stationary world, namely that variables should be retained if their noncentralities exceed unity, still applies in settings with structural breaks. This provides support for model selection at looser than conventional settings, albeit with many additional features explaining the forecast performance, and with the caveat that retaining irrelevant variables that are subject to location shifts can worsen forecast performance.

Keywords: model selection; forecasting; location shifts; significance level; Autometrics

Citation: Castle, Jennifer L., Jurgen A. Doornik, and David F. Hendry. 2021. Selecting a Model for Forecasting. *Econometrics* 9: 26. https://doi.org/10.3390/econometrics9030026

Academic Editor: Neil Ericsson

Received: 9 November 2018
Accepted: 17 June 2021
Published: 25 June 2021

1. Introduction

There are many approaches to formulating models when the sole objective is forecasting, from the very parsimonious through to large systems. However, there is little agreement on which performs best on a forecasting criterion: see Makridakis and Hibon (2000) and Fildes and Ord (2002) for evidence from forecast competitions. Clements and Hendry (2001) suggest that this lack of agreement is the result of intermittent distributional shifts that affect alternative formulations in different ways. We address this puzzle by analysing the selection of models in the pursuit of optimal mean square forecast error (MSFE) in settings with structural breaks.[1]

We focus on regression models that are linear in the parameters, and consider model selection that is controlled by the nominal significance level for statistical significance. Loose significance levels have been shown to be optimal to select regression models for stationary processes if evaluating on a one-step-ahead MSFE. Shibata (1980) showed that the Akaike information criterion (AIC, see Akaike 1973) is an asymptotically efficient selection method when the data generating process (DGP) is an infinite-order process; also see Ing and Wei (2003). Many other criteria have been proposed that aim to have optimal properties in certain settings but information criteria alone are not a sufficient principle for selecting models as they do not ensure congruence, so a misspecified model could be selected: see Bontemps and Mizon (2003). We explore general-to-specific (Gets) model selection in the simulation exercise to narrow down the class of forecasting models to undominated models. This yields well-specified encompassing models in sample, albeit nonstationarities may preclude those benefits continuing over the forecast horizon.

The theoretical analysis commences with a bivariate conditional model that is part of a three-variable system in which the selection decision is whether to retain or exclude one of the regressors. This is empirically relevant as demonstrated by UK inflation, where autoregressive (AR) forecasting models are augmented with the unemployment rate. The bivariate model is analysed first in a stationary setting. This is extended to a nonstationary

settings where location shifts occur at or near the forecast origin. The static setting still requires forecasts of the conditioning variables, and alternative forecasting devices are considered, including the two extremes of the class of robust forecasting devices proposed by Castle et al. (2015), the sample mean and the random walk. The results confirm that regressors should be retained for forecasting if their noncentralities exceed unity, regardless of whether or not there is a structural break, or of the forecasting device used. These analytic results map to a selection significance level of 16% in the bivariate case, much looser than conventional significance levels used. The results closely match that of AIC, which can be interpreted as a likelihood ratio χ^2 test for a pair of nested models with one degree of freedom and a penalty of two, and also gives a significance level of approximately 16%: see Pötscher (1991) and Leeb and Pötscher (2009).

A key source of forecast failure is an induced shift in the equilibrium mean of the variable being forecast, irrespective of whether those conditioning variables are included in the forecasting model; see the taxonomy in Hendry and Mizon (2012). Consequently, the simulation exercise evaluates a wide range of settings including larger models, break types and magnitudes at or near the forecast origin, and the method of forecasting. We consider a range of significance levels from the very tight (0.001), eliminating almost all potentially irrelevant variables, to the very loose (0.50), enabling retention of relevant variables even if they are only marginally significant. The results enable evaluation of the costs when either omitting relevant variables, or from incorrectly retaining irrelevant variables. Overall, the results support looser than conventional significance levels for selecting forecasting models, with a 10% target significance level often producing superior forecasts.

This paper is structured as follows. Section 2 outlines the aims of this paper, then Section 3 formulates the model framework that is analysed. Section 4 considers the choice of selection significance level for forecasting in a stationary DGP. Section 5 analyses selection in a nonstationary DGP where a location shift occurs out of sample in one of the regressors, and investigates the consequences of that variable's inclusion or exclusion in the forecasting model. Section 6 considers the impacts on selection of in-sample shifts using different forecasting devices. The analytic results are summarized in Section 7. Sections 8 and 9 present simulation design and evidence on the performance of the various approaches, examining the preferred significance level to minimize MSFE across experimental designs. Section 10 concludes this paper. Appendix A provides analytical calculations and Supplementary Tables are given in Appendix B.

2. Empirical Motivation

An empirical example of inflation forecasting motivates our interest in structural breaks and their roles in forecast accuracy and the selection of regressors. Two popular models within this large literature include single-equation forecasting models based on past inflation and so-called 'Phillips curve forecasts'. The former usually consist of univariate models such as autoregressive integrated moving average (ARIMA) models. In the latter, the univariate model is augmented with an activity variable such as the unemployment rate or output gap; see Stock and Watson (2009).

The framework considered below, although static, can be applied to these two models where the econometrician wishes to determine whether to augment a univariate forecasting model with the contemporaneous unemployment rate. This 'exogenous' variable is subject to breaks in the form of location shifts, which may occur at or near the forecast horizon. Figure 1 records[2] the quarterly observations on the annual percentage inflation in UK consumer price index, π_t, and the UK unemployment rate as a percentage, U_t, along with a broken mean obtained by step indicator saturation (SIS, see Castle et al. 2015) at a nominal significance level $\alpha = 0.1\%$.

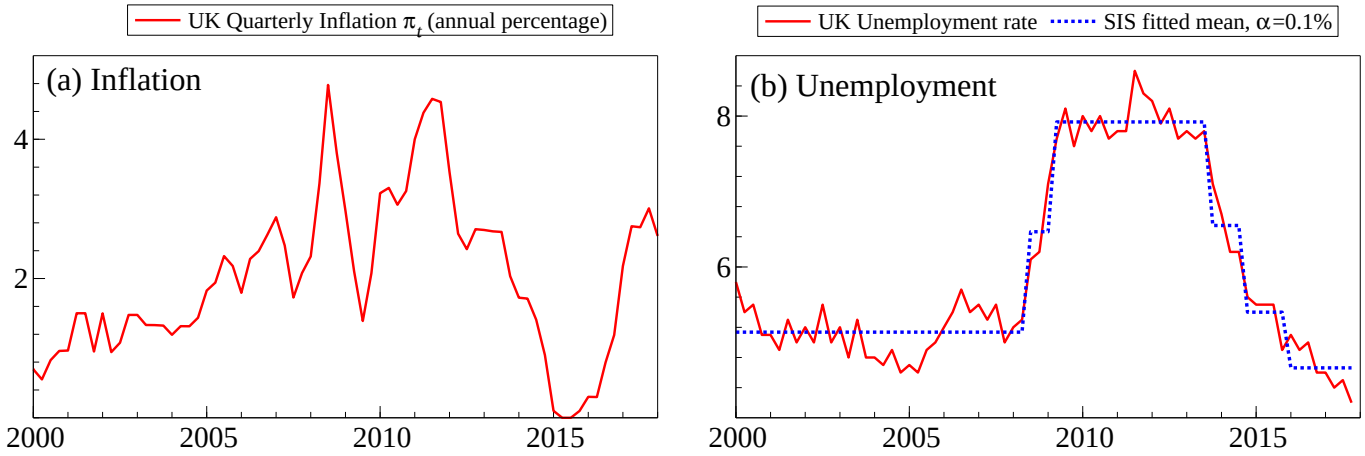

Figure 1. (**a**) Quarterly average of CPI 12 month inflation rates for the UK (percent per annum); (**b**) quarterly UK unemployment rate in percent, with SIS detected mean shifts at $\alpha = 0.1\%$.

The analytics derived below correspond to a Phillips curve formulation (model M_1), a univariate AR model (M_2) and selection applied to the unemployment rate using a significance level of 0.16 (M_3). Using model-specific coefficients μ, β_i, γ_i and error term ν_i, the three models are:

$$\begin{aligned}
M_1 &: \Delta\pi_t = \mu + \textstyle\sum_{i=1}^{4} \beta_i \Delta\pi_{t-i} + \sum_{i=0}^{4} \gamma_i U_{t-i} + \nu_t, \\
M_2 &: \Delta\pi_t = \mu + \textstyle\sum_{i=1}^{4} \beta_i \Delta\pi_{t-i} + \nu_t, \\
M_3 &: \Delta\pi_t = \mu + \textstyle\sum_{i=1}^{4} \beta_i \Delta\pi_{t-i} + \sum_{i=0}^{4} \gamma_i^* U_{t-i} + \nu_t,
\end{aligned}$$

where $\Delta\pi_t = \pi_t - \pi_{t-1}$. Selection using Autometrics at $\alpha = 0.16$ is denoted by *, e.g., $\gamma_0^* = 0$ implies that the contemporaneous unemployment rate is not selected. Dynamics are included to account for any autocorrelation. The forecasting models are estimated over the period 2000Q1–2013Q4, producing one-quarter-ahead inflation forecasts for the period 2014Q1–2017Q4 evaluated on MSFE. Selection at 16% results in U_{t-1} being retained, with a p-value of 0.149, so would not be retained under a commonly used 5% significance level. Longer lags of the unemployment rate were not retained.

Table 1 reports the square root of the MSFEs (RMSFE) for one-step-ahead forecasts over the sample that was held back. Three cases are considered corresponding to the analytics below: (a) known U_t, (b) forecast $\widehat{U}_t$ using the in-sample mean, and (c) forecast $\widehat{U}_t$ using U_{t-1}. Method (a) is infeasible; method (c) is the random walk forecast. When U_t is known, model M_3 outperforms M_1 and M_2, although the differences are not statistically significant. As this is infeasible, the random walk forecast combined with selection matches the RMSFE of knowing U_t. This shows that selection can be beneficial. The next four sections formalize the framework to establish the optimal significance level for selection.

Table 1. Root mean square error of one-step forecast for $\Delta\pi_t$ over the period 2014Q1–2017Q4.

Conditioning on	**M_1**	**M_2**	**M_3**
Known U_t	0.535	0.530	0.515
Mean forecast for U_t	0.519	0.530	0.542
Random walk forecast for U_t	0.549	0.530	0.515

3. The Analytic Design

In this section, we specify the analytic design, consisting of a three-variable DGP and two different models for that DGP. In later sections, we introduce a third model that involves selection. Together, these mimic the models M_1, M_2, and M_3 that were introduced above.

The DGP is a static vector autoregression (VAR) for variables y, x_1, x_2 with coefficients β_i, μ_i and error terms ϵ, η_1, η_2 structured as:

$$\begin{pmatrix} 1 & -\beta_1 & -\beta_2 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} y_t \\ x_{1,t} \\ x_{2,t} \end{pmatrix} = \begin{pmatrix} \beta_0 \\ \mu_1 \\ \mu_2 \end{pmatrix} + \begin{pmatrix} \epsilon_t \\ \eta_{1,t} \\ \eta_{2,t} \end{pmatrix}. \tag{1}$$

Using $\mathbf{y}_t' = (y_t : x_{1,t} : x_{2,t})$ and $\boldsymbol{\mu}' = (\mu_y : \mu_1 : \mu_2)$, assuming normality, we can write (1) as:

$$\mathbf{y}_t \sim \mathrm{IN}_3[\boldsymbol{\mu}, \boldsymbol{\Sigma}]. \tag{2}$$

IN_3 denotes a three-dimensional independent normal distribution, here with mean $\boldsymbol{\mu}$ and variance $\boldsymbol{\Sigma}$. Without loss of generality we set the variance of x_1 and x_2 to one, $\mathrm{V}[x_{i,t}] = \sigma_{ii}^2 = 1$, and the correlation between x_1 and x_2 to ρ:

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_\epsilon^2 & 0 & 0 \\ 0 & 1 & \rho \\ 0 & \rho & 1 \end{pmatrix}. \tag{3}$$

Unless otherwise noted, Figures 2–8 use the following parameter values in calculations: $\beta_0 = 5$, $\beta_1 = 1$, $\sigma_\epsilon^2 = 1$, $\mu_1 = \mu_2 = 2$, $\rho = 0.5$, $M = 10^5$, $T = 50$ and (when there is a break in μ_2) $\delta = 4$.

Although a static DGP may seem restrictive, the main role of adding dynamics to this three-variable VAR would be to slow adjustments to location shifts. Such dynamics are considered in the simulation exercise in Section 9. The analytic design ensures the assumptions required for valid application of a single t-test are satisfied. In practice, selection from a carefully designed general model including long lags and saturation estimators should deliver approximately martingale-difference normal residuals. While it may be more intuitive to lag the exogenous regressors in the DGP for forecasting purposes, none of the results would change. The current set up naturally leads to analyses of the forecasting models for the contemporaneous exogenous regressors, allowing a comparison of alternative devices and an assessment of open models, see Hendry and Mizon (2012).

Throughout, we assume that the sampling variation of estimates of μ_i can be neglected, and use the population values to focus on the impacts of location shifts. Then (1) implies $\mathrm{E}[y_t] = \mu_y = \beta_0 + \beta_1\mu_1 + \beta_2\mu_2$ with:

$$y_t = \mu_y + \beta_1(x_{1,t} - \mu_1) + \beta_2(x_{2,t} - \mu_2) + \epsilon_t. \tag{4}$$

Considering the conditional model (4), we compare M_1, which includes both weakly exogenous regressors, and M_2, which excludes x_2:

$$\mathrm{M}_1 : y_t = \beta_0 + \beta_1 x_{1,t} + \beta_2 x_{2,t} + \epsilon_t, \tag{5}$$

$$\mathrm{M}_2 : y_t = \phi_0 + \gamma_1 x_{1,t} + \nu_t, \tag{6}$$

where Appendix A.1 summarises ϕ_0, γ_1, ν_t and σ_ν^2.

The choice between M_1 and M_2 will depend on a test of significance of $x_{2,t}$. The usual Student's t-statistic for β_2 is

$$\mathrm{t}_\beta = \frac{\widehat{\beta}_2}{s.e.\left(\widehat{\beta}_2\right)} \sim \mathrm{t}\left(T-k, \psi_\beta\right),$$

where $\mathrm{t}\left(T-k, \psi_\beta\right)$ indicates a singly noncentral Student's t-distribution with ψ_β nonzero under the alternative hypothesis. Here, $T-k$ is the degrees of freedom, and

$$\psi_\beta^2 = \frac{T\beta_2^2(1-\rho^2)}{\sigma_\epsilon^2} \tag{7}$$

is the squared noncentrality parameter under the alternative.

4. Selection in a Stationary DGP

We start by analysing the forecast errors of the two models that were introduced, denoted M_1 and M_2, in the absence of breaks. The analysis is then augmented in Section 4.2 by introducing selection of regressors in M_3, and the influence of the significance level on the selection decision in Section 4.3. In this section, we assume that there are no breaks in the DGP.

4.1. Known Future Values of Regressors

The one-step-ahead forecast errors from M_1 are denoted $\widehat{\epsilon}$ and those from M_2 $\widetilde{\epsilon}$. The mean square forecast errors are written as MSFE_1 and MSFE_2 respectively. We look at the conditions for $\text{MSFE}_2 \leq \text{MSFE}_1$. An estimated intercept is always retained which maintains comparability between M_1 and M_2.

When there are no breaks, the parameter estimates for M_1 are unbiased, $\mathsf{E}[\widehat{\epsilon}_{T+1|T}] = 0$, so:

$$\text{MSFE}_1 = \mathsf{E}\left[\widehat{\epsilon}^2_{T+1|T}\right] = \sigma^2_\epsilon\left(1 + \frac{3}{T}\right), \tag{8}$$

which is the unconditional MSFE formula for the impact of estimating 3 parameters under the assumption of correct model specification. For M_2, despite the misspecification when $\beta_2 \neq 0$, $\mathsf{E}[\widetilde{\epsilon}_{T+1|T}] = 0$ and the mean square forecast error is:

$$\text{MSFE}_2 = \mathsf{E}\left[\widetilde{\epsilon}^2_{T+1|T}\right] = \sigma^2_\nu\left(1 + \frac{2}{T}\right), \tag{9}$$

where $\sigma^2_\nu = \sigma^2_\epsilon\left(1 + T^{-1}\psi^2_\beta\right) \geqq \sigma^2_\epsilon$. There is one less parameter to estimate, traded off against a larger equation variance (see Appendix A.2 for derivations).

If the objective is to minimize MSFE, M_2 should be used to forecast when $\text{MSFE}_2 \leq \text{MSFE}_1$, which requires:

$$\sigma^2_\nu\left(1 + \frac{2}{T}\right) \leq \sigma^2_\epsilon\left(1 + \frac{3}{T}\right). \tag{10}$$

From (7), this occurs when $\psi^2_\beta \leq T/(T+2)$.

Figure 2 records the one-step-ahead values of MSFE_1 and MSFE_2 for known $x_{i,T+1}$, $i = 1, 2$, for the DGP given by (1) and (2). We let β_2 vary along the horizontal axis to get a range of noncentralities in the set $\psi_\beta = [0, 4]$ using (7).

The results confirm that x_2 should be retained if its noncentrality exceeds approximately 1. The result converges to 1 as $T \to \infty$, because the information content of the regressor outweighs the parameter estimation cost for one-step forecasts, regardless of the correlation between x_1 and x_2.

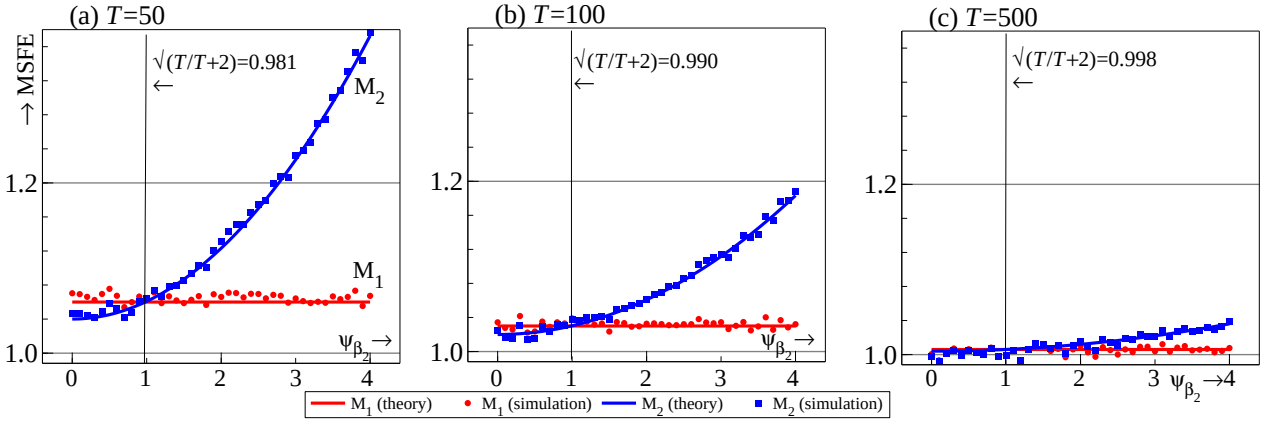

Figure 2. MSFE_1 (solid lines computed from (8), circles by simulation) and MSFE_2 (dashed line computed from (9), squares by simulation).

4.2. Selecting Regressors

Although M_1 and M_2 provide the extremes of always/never retaining x_2, in practice, selection will be applied. From (5), $x_{2,t}$ will be omitted if $\mathrm{t}^2_{\beta_2=0} < c^2_\alpha$. Using the approximation that:

$$\mathrm{t}_{\beta_2=0} = \frac{\widehat{\beta}_2}{s.e.\left[\widehat{\beta}_2\right]} \approx \frac{\sqrt{T(1-\rho^2)}\widehat{\beta}_2}{\sigma_\epsilon},$$

implies:

$$\widehat{\beta}^2_2 < \frac{c^2_\alpha \sigma^2_\epsilon}{T(1-\rho^2)}. \tag{11}$$

Thus, retention of $x_{2,t}$ will depend on α and ψ^2_β for a given draw.

Forecasts in repeated sampling will be based on a mixture of M_1 and M_2 depending on whether $x_{2,t}$ is retained in each draw. The MSFE of the selected model, called M_3, will be a weighted average of the MSFEs of M_1 and M_2, with the weights given by the probability that $x_{2,t}$ is retained:

$$\begin{aligned}
\mathrm{MSFE}_3 &= \mathrm{p}_\alpha(\psi_\beta)\mathrm{MSFE}_1 + \left(1-\mathrm{p}_\alpha(\psi_\beta)\right)\mathrm{MSFE}_2 \\
&= \mathrm{MSFE}_1 + \left(1-\mathrm{p}_\alpha(\psi_\beta)\right)\left(\mathrm{MSFE}_2 - \mathrm{MSFE}_1\right) \qquad (12)\\
&\approx \mathrm{MSFE}_1 + \sigma^2_\epsilon T^{-1}\left(1-\mathrm{p}_\alpha(\psi_\beta)\right)\left(\psi^2_\beta - 1\right), \qquad (13)
\end{aligned}$$

where ψ^2_β is given by (7), with:

$$\mathrm{p}_\alpha(\psi_\beta) = \Pr\left(\mathrm{t}^2_{\beta_2=0} \geqq c^2_\alpha\right).$$

From the last term in (13), it is clear that $\mathrm{MSFE}_3 \leq \mathrm{MSFE}_1$ whenever $\psi^2_\beta \leq 1$. Moreover, $\mathrm{p}_\alpha(\psi_\beta)$ will be low when $\psi^2_\beta \leq 1$, so M_2 will usually be selected. Note that $\mathrm{p}_\alpha(\psi_\beta) = \alpha$ when $\beta_2 = 0$. However, MSFE_3 is a highly nonlinear function of ψ^2_β entering directly and indirectly, as well as of α which also influences $\mathrm{p}_\alpha(\psi_\beta)$ nonlinearly.

Figure 3 records the ratio of MSFE_3 to MSFE_1, for a range of ψ^2_β, which from (13) is given by:

$$\frac{\mathrm{MSFE}_3}{\mathrm{MSFE}_1} \approx 1 + (T+3)^{-1}\left(1-\mathrm{p}_\alpha(\psi_\beta)\right)\left(\psi^2_\beta - 1\right). \tag{14}$$

Selection delivers a 1.8% improvement in MSFE relative to M_1 under the null when $\psi^2_\beta = 0$ with $\alpha = 0.05$ or tighter, but for looser α, e.g., at 0.5, $\mathrm{p}_\alpha(\psi_\beta) = 0.5$ when $x_{2,t}$ is irrelevant so the benefits of selection are halved. Selection is most costly at intermediate noncentralities

under the alternative, where, e.g., the largest increase in MSFE relative to M_1 is 3% at $\alpha = 0.05$ for $T = 50$, but is over 9% for $\alpha = 0.001$ at its peak. The hump shape reflects the nonlinear trade-off as the noncentrality of $x_{2,t}$ increases from the cost of omitting $x_{2,t}$ rising as its signal is stronger, but the probability of retaining $x_{2,t}$ also increases. While the magnitude of the maximal loss may seem small for intermediate values of α, this example considers the selection of just one regressor. In practice, selection is applied when there are multiple potential regressors, and the loss associated with selection at a given significance level is cumulated across all potential regressors, as seen in the simulation results below.

The selection rule that $x_{2,t}$ should be retained if $\psi_\beta^2 > 1$ is evident $\forall \alpha$, but unfortunately the forecaster does not know ψ_β^2. If it was known, the optimal α is 0 for $\psi_\beta^2 < 1$ and 1 for $\psi_\beta^2 > 1$. We next look at the choice of α to minimize cost in terms of improvements in MSFEs for an unknown ψ_β^2.

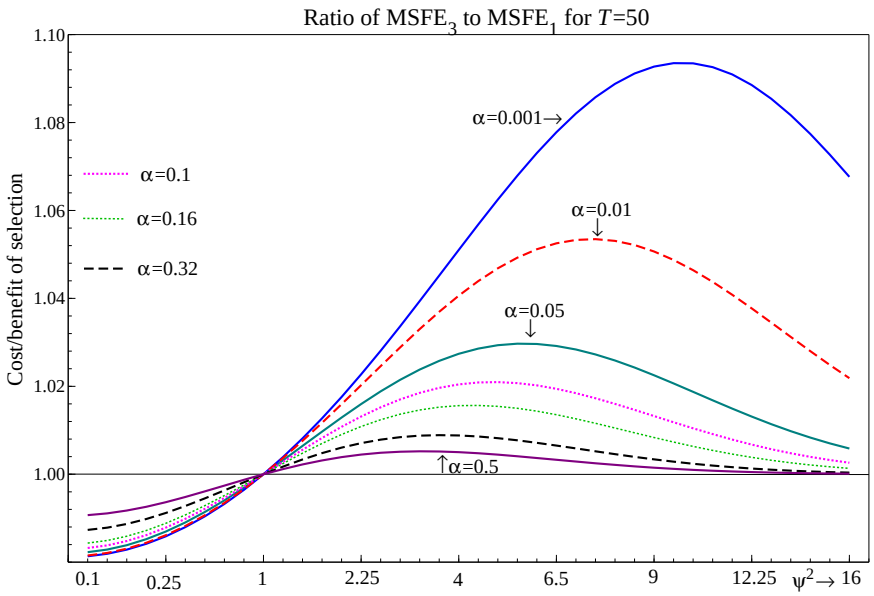

Figure 3. The costs/benefits of selection measured by $\frac{\text{MSFE}_3}{\text{MSFE}_1}$ in (14).

4.3. The Choice of Significance Level

Equation (11) must hold for x_2 to be excluded at the chosen significance level. On average, that inequality requires:

$$\mathrm{E}\left[\widehat{\beta}_2^2\right] = \mathrm{V}\left[\widehat{\beta}_2\right] + \beta_2^2 = \beta_2^2 + \frac{\sigma_\epsilon^2}{T(1-\rho^2)} < \frac{c_\alpha^2 \sigma_\epsilon^2}{T(1-\rho^2)},$$

assuming unbiasedness. Equating that inequality for β_2^2 with $\psi_\beta^2 < 1$ from (10) gives the boundary for the critical value c_α in which selection results in a smaller MSFE due to the omission–estimation trade-off:

$$\beta_2^2 = \frac{\sigma_\epsilon^2 (c_\alpha^2 - 1)}{T(1-\rho^2)} \leq \frac{\sigma_\epsilon^2}{T(1-\rho^2)}.$$

This implies that $c_\alpha^2 = 2$ at the boundary, or an approximate significance level of $\alpha = 0.16$.

The theoretical probability of retaining x_2 for $\beta_2 > 0$ at $\alpha = 0.16$ using $\mathrm{E}[\mathrm{t}_{\widehat{\beta}_2}] = \psi_\beta$ is:

$$\Pr\left(\mathrm{t}_{\widehat{\beta}_2} \geqq c_\alpha\right) = \Pr\left(\mathrm{t}_{\widehat{\beta}_2} - \psi_\beta \geqq c_\alpha - \psi_\beta\right).$$

This gives the retention probabilities recorded in Table 2.

These results are close to the implied significance level for the AIC in Campos et al. (2003). This can have a cumulative effect, as shown in Figure 4 which records values of the term $(1 - \mathrm{p}_\alpha(\psi_\beta))$ where there are five independent regressors, all with the same ψ^2_β. The probability of retaining all five variables is low even at loose significance levels unless the noncentralities are large. At $\psi^2_\beta = 9$ the gap between $\alpha = 0.05$ and $\alpha = 0.16$ is 29%, demonstrating large benefits to a looser significance level for the retention of relevant regressors. The trade-off is that more irrelevant variables will be retained, and this can be costly if those variables are subject to breaks, which we next explore.

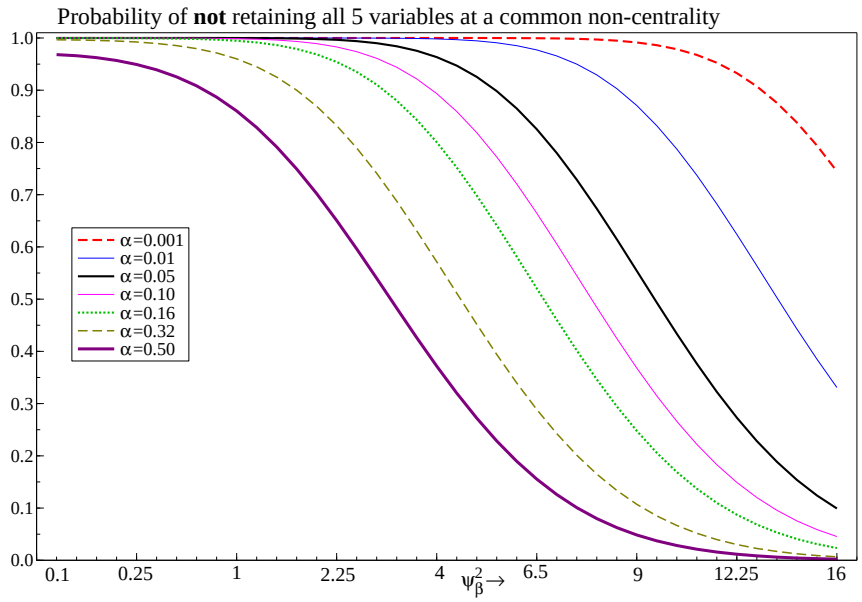

Figure 4. Values of $(1 - \mathrm{p}_\alpha(\psi_\beta))$ for five independent regressors with the same noncentrality for a range of α and ψ^2_β.

Table 2. Retention probabilities for individual t-tests given $\mathrm{E}[\mathrm{t}_{\hat{\beta}_2}] = \psi_\beta$.

ψ_β	1	2	3	4
$\mathrm{P}_{0.16}$	0.34	0.72	0.94	0.995
$\mathrm{P}_{0.05}$	0.16	0.51	0.85	0.98

5. An Out-of-Sample Shift in the Regressors

The analysis of the previous section is augmented by the introduction of a break in Section 5.1. This break is immediately after the estimation sample, while in Section 6 it is applied to the last in-sample observation. We distinguish between whether the future values of the regressors are known (Section 5.2) or unknown (Section 5.4). The role of selection is studied again (Section 5.3), and we look at the random walk as a device to forecast future values of the regressors in Section 5.5. Forecasting devices based on full in-sample information and a random walk are the extremes of the class in Castle et al. (2015), but there is no information in sample regarding the break to help either device.

5.1. Specification of the Out-of-Sample Shift

Consider a mean shift of size δ in x_2 at $T+1$ with the forecast origin at T, so the shift coincides with the one-step-ahead forecast. The DGP has the same structure as (1)–(3) with the parameters $(\beta_1\ \beta_2)$ of the conditional model constant:

$$\begin{aligned} x_{1,t} &= \mu_1 + \eta_{1,t} && t = 1,\dots,T+1, \\ x_{2,t} &= \begin{cases} \mu_2 + \eta_{2,t} & t = 1,\dots,T, \\ \mu_2 + \delta + \eta_{2,t} & t = T+1. \end{cases} \end{aligned} \tag{15}$$

Since (15) entails:

$$\begin{aligned} y_{T+1} &= \beta_0 + \beta_1 x_{1,T+1} + \beta_2 x_{2,T+1} + \epsilon_{T+1} \\ &= (\mu_y + \beta_2\delta) + \beta_1(x_{1,T+1} - \mu_1) + \beta_2(x_{2,T+1} - \mu_2 - \delta) + \epsilon_{T+1}, \end{aligned} \tag{16}$$

then $\beta_2\delta \neq 0$ induces a location shift in the relationship between y_{T+1} and its in-sample determinants unless the future $x_{2,T+1}$ is known at time T. As shown in all forecast-error taxonomies (see e.g., Clements and Hendry 1998), shifts in the equilibrium mean are the most pernicious source of forecast failure, whereas changes in the parameters of mean-zero variables have only a variance impact. Omitting $x_{2,T+1}$ from (16) as in M_2 will create the same location shift. Thus, there is little loss of generality by only considering shifts in the regressors.

We first evaluate the trade-off to omitting $x_{2,t}$ for known future exogenous regressors, emulating the above results as the break which occurs in the forecast period is modeled in the known $x_{2,T+1}$.

5.2. Known Future Values of Regressors

The one-step-ahead forecasts for M_1 given (15), in which values of $\mathbf{x}_{T+1}$ are assumed to be known at T, are unbiased when the parameter estimates are unbiased. The mean square forecast error of M_1 (see Appendix A.3 for derivations) is:

$$\mathrm{MSFE}_1 = \mathrm{E}\left[\widetilde{\widehat{\epsilon}}^2_{T+1|T+1}\right] = \sigma^2_\epsilon\left(1 + \frac{1}{T(1-\rho^2)}\left(\delta^2 + 2 - \rho\right)\right), \tag{17}$$

which does not depend on ψ^2_β. Comparison with (8) highlights the effects of the location shift: δ^2 enters the MSFE despite the shift being 'known' given $x_{2,T+1}$, and MSFE_1 is no longer independent of ρ. (17) also reveals the additional costs of including an irrelevant regressor which shifts out of sample as δ^2 enters even when $\beta_2 = 0$, although it is scaled by $T(1-\rho^2)$ so larger samples mitigate its effect.

For M_2 (which omits the regressor $x_{2,t}$), the expectation of the forecast error is $\mathrm{E}[\widetilde{\widehat{\epsilon}}_{T+1|T+1}] = \beta_2\delta$, so the forecasts are biased by the shift in the omitted variable. The one-step-ahead MSFE for M_2 is:

$$\mathrm{MSFE}_2 = \mathrm{E}\left[\widetilde{\widehat{\epsilon}}^2_{T+1|T+1}\right] = \sigma^2_\epsilon + \beta^2_2\left(1 - \rho^2 + \delta^2\right) + 2T^{-1}\sigma^2_\epsilon\left(1 + T^{-1}\psi^2_\beta\right), \tag{18}$$

where β^2_2 enters directly so the MSFE is a function of ψ^2_β, unlike for M_1. Comparison with (9) reveals the role that ρ and δ^2 play. When $\beta_2 = 0$, so M_2 is the correct model, (18) collapses to (9).

Assuming a criterion of minimizing one-step-ahead MSFE, using (10), $\mathrm{MSFE}_2 \leq \mathrm{MSFE}_1$ requires:

$$\delta^2\left(\psi^2_\beta - 1\right) + \psi^2_\beta\left(1 - \rho^2\right)\left(1 + 2T^{-1}\right) - \rho < 0, \tag{19}$$

which depends on estimation uncertainty and therefore does not simplify neatly. However, the solution is close to 1 for reasonable values of ρ. For example, when $\rho = 0.5$, $T = 50$ and $\delta = 4$, then $\psi^2_\beta < 0.983$, or $|\psi_\beta| < 0.991$, results in a smaller MSFE_2 compared to MSFE_1.

Figure 5 demonstrates the close approximation to a trade-off at $\psi_\beta = 1$ which holds regardless of the break. Thus, even knowing there is a shift in x_2 does not affect the choice of forecasting model between including or omitting x_2: always (never) include for $\psi_\beta^2 \geq 1$ ($\psi_\beta^2 < 1$).

5.3. Selecting Regressors

Following Section 4.2, a t-test for statistical significance will be conducted on $x_{2,t}$ in sample and a decision to retain or exclude $x_{2,t}$ will be made at c_α for a given draw. Hence, MSFE_3 will be a weighted average of MSFE_1 and MSFE_2, using (12):

$$\text{MSFE}_3 = \text{MSFE}_1 + (1 - \text{p}_\alpha(\psi_\beta)) \left(\sigma_\epsilon^2 T^{-1} \left[\psi_\beta^2 \left\{ 1 + \frac{\delta^2}{(1-\rho^2)} \right\} - \frac{\delta^2 + 2 - \rho}{(1-\rho^2)} \right] \right). \tag{20}$$

The term in square brackets is scaled by T^{-1}. As before, the difference between MSFE_1 and MSFE_3 diminishes as the sample size increases. When $\psi_\beta^2 = 0$, the first term in square brackets in (20) drops out and the benefits of selection relative to MSFE_1 are evident as the second term must be negative. The magnitude of δ^2 affects both MSFE_1 and MSFE_2 but, from (20), the first δ^2 term is multiplied by ψ_β^2 whereas the second offsetting term is not, so the effect of the location shift is exacerbated if $\psi_\beta^2 > 1$.

Figure 5 compares the MSFEs of M_1 from (17), M_2 from (18), and M_3 using (20) at three illustrative values of α. The profiles of the MSFEs mirror the analytical results for the no break case. Selection outperforms the estimated DGP for $\psi_\beta^2 < 1$ despite a break, and remains close to the MSFE_1 at $\alpha = 0.16$ for $\psi_\beta^2 > 1$.

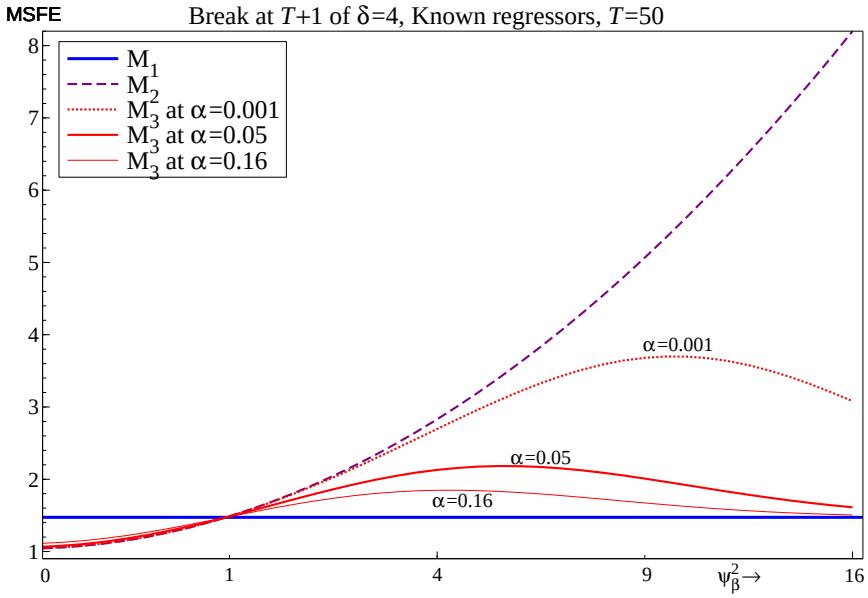

Figure 5. MSFE comparisons of M_1, M_2 and M_3 at 3 illustrative values of α for known future exogenous regressors where the break occurs in the mean of x_2 at $T+1$.

5.4. Unknown Future Values of Regressors

Now consider when the future values of the regressors are unknown. We use two devices to obtain forecasts of $x_{i,T+1}$, $i = 1, 2$: the in-sample mean or a random walk. The random walk is biased for unanticipated location shifts but does not result in systematic bias following a location shift, whereas the in-sample mean is persistently biased following a location shift unless updated. The two devices comprise the two extremes of using either the full in-sample data or only the last observation to produce the forecasts of the weakly exogenous regressors.[3]

Although the link between y and the x_i stays constant, forecasts when the $x_{i,T+1}$ are unknown will fail if the shift at $T+1$ is not anticipated, inducing a shift in y_{T+1}. This will lead to forecast failure as the in-sample mean μ_y shifts to $(\mu_y + \beta_2\delta)$ at $T+1$ but would be forecast to be μ_y.

The forecasts based on in-sample estimates from (15) when μ_1 and μ_2 are not zero are given by:

$$\overline{x}_{1,T+1|T} = \widehat{\mu}_1 = \tfrac{1}{T}\textstyle\sum_{t=1}^{T} x_{1,t} = \mu_1 + \overline{\eta}_1, \tag{21}$$

$$\overline{x}_{2,T+1|T} = \widehat{\mu}_2 = \tfrac{1}{T}\textstyle\sum_{t=1}^{T} x_{2,t} = \mu_2 + \overline{\eta}_2, \tag{22}$$

so will miss the unknown break. When the break occurs in x_2, the MSFEs will worsen for $\beta_2 \neq 0$. As before, we consider the sampling variation in estimating the means as small compared to the impact of shifts, so we approximate by taking T sufficiently large that $\widehat{\mu}_i \approx \mu_i$.

Replacing the unknown $x_{i,T+1}$ by μ_i leads to forecasting y_{T+1} by the in-sample mean for both M_1 and M_2, see Appendix A.4. Both face the same forecast bias, $\mathrm{E}[\widehat{\widetilde{\epsilon}}_{T+1|T}] = \mathrm{E}[\widetilde{\widetilde{\epsilon}}_{T+1|T}] = \beta_2\delta$ which is the same bias as M_2 with known regressors. Parameter estimation adds terms of $\mathrm{O}_p\left(T^{-1}\right)$. Hence, ignoring $\mathrm{O}_p\left(T^{-1}\right)$ terms, $\mathrm{MSFE}_1 = \mathrm{MSFE}_2$:

$$\mathrm{E}\left[\widehat{\widetilde{\epsilon}}^2_{T+1|T}\right] = \mathrm{E}\left[\widetilde{\widetilde{\epsilon}}^2_{T+1|T}\right] = \beta_2^2\delta^2 + \sigma_\epsilon^2 + \left(\beta_1^2 + \beta_2^2 + 2\rho\beta_1\beta_2\right). \tag{23}$$

When $\beta_2 = 0$, the MSFE is $\sigma_\epsilon^2 + \beta_1^2$, so is inflated relative to the known regressors case as $x_{1,T+1}$ must also be forecast. However, the in-sample mean forecast is the best forecast device for $x_{1,T+1}$ in this setting (in terms of minimum MSFE) as $x_{1,T+1}$ is stationary and not subject to a location shift. Selection will have little or no noticeable impact when $\mathrm{MSFE}_2 \approx \mathrm{MSFE}_1$, as this will also result in $\mathrm{MSFE}_3 \approx \mathrm{MSFE}_1$.

Figure 6 records the MSFEs for M_1 and M_2 when there is a break in x_2 at $T+1$, comparing known and unknown regressors using the in-sample mean to forecast $x_{i,T+1}$, $i = 1, 2$ in the unknown regressor case, i.e., the figure records (17), (18) and (23), (solid/dashed/dotted lines). Simulation outcomes are checked to capture $\mathrm{O}_p\left(T^{-1}\right)$ effects but they are negligible so are not recorded. Figure 6 includes the random walk forecasts and the M_1 and M_2 results for the known regressor case are repeated from Figure 5 to facilitate comparison.

The simulation outcomes where parameters are estimated closely match the analytic results. For known regressors for MSFE_1, the break in μ_2 does not affect the MSFE as it is captured in $x_{2,T+1}$: even at $\delta = 4$ for $T = 100$, $\mathrm{MSFE}_1 = 1.23$ for the parameters given in the figure which is only slightly greater than σ_ϵ^2. However, when $\mathbf{x}_{T+1}$ is unknown both M_1 and M_2 are affected by the break in $x_{2,T+1}$. Simulation outcomes again closely match the theory for the unknown break case, and show that the choice of whether to retain or exclude $x_{2,t}$ is not important in a forecasting context. The unanticipated break dominates any forecast error resulting from model misspecification. Increasing the sample size does mitigate the MSFE costs but the MSFE premium relative to known regressors is maintained for all ψ_β^2. Increasing the number of relevant exogenous regressors that shift will increase the MSFE at $\psi_\beta^2 = 0$, shifting the MSFE trajectories up.

These results show that in this static setting of location shifts, if the break occurs in the forecast period and is unknown and unpredictable, then the retention of x_2 is irrelevant (other than parameter estimation uncertainty), as neither M_1 nor M_2 capture the shift which dominates the MSFE. *Parsimony, or lack thereof, neither helps nor hinders much in this setting*. Moreover, selection does not substantively affect the outcome as $\mathrm{MSFE}_3 \approx \mathrm{MSFE}_1$.

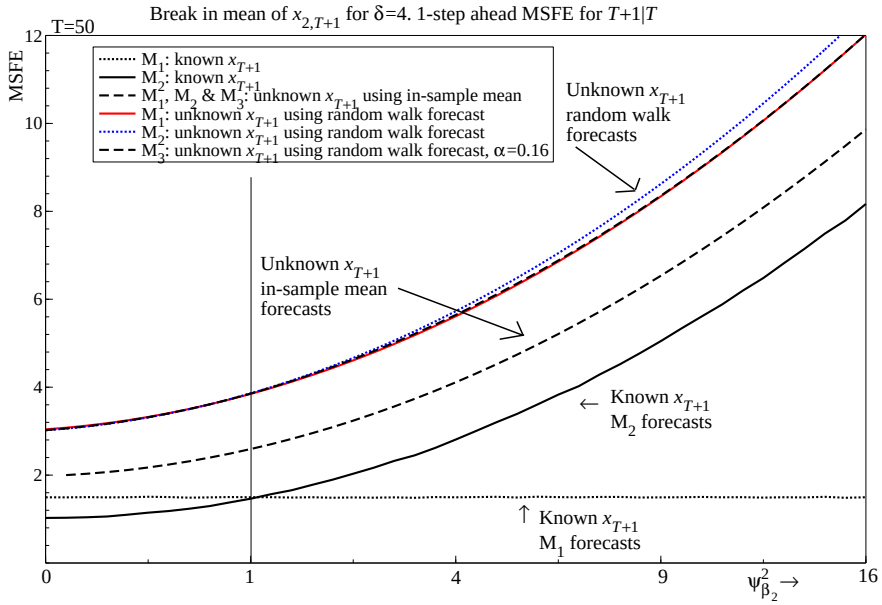

Figure 6. MSFE comparisons between M_1, M_2 and M_3 for known and unknown future exogenous regressors including in-sample mean and random walk forecasts, where the break occurs in the mean of x_2 at $T+1$.

5.5. Forecasting Regressors with a Random Walk

We now consider using a random walk to forecast the exogenous variables:

$$\overline{\overline{x}}_{1,T+1|T} = x_{1,T}, \tag{24}$$

$$\overline{\overline{x}}_{2,T+1|T} = x_{2,T}. \tag{25}$$

Such a device is not robust in this setting as the forecasts are made before the shift, and robustness refers to forecasting properties that are insensitive to a feature in the DGP, such as after a location shift.

Although the last in-sample observation is an imprecise measure of the out-of-sample mean, it is unbiased when there are no location shifts (as there are no dynamics in the DGP), so $\mathrm{E}[x_{1,T}] = \mu_1$ and $\mathrm{E}[x_{2,T}] = \mu_2$, and hence $\mathrm{E}[\Delta x_{1,T+1}] = 0$ and $\mathrm{E}[\Delta x_{2,T+1}] = \delta$.

The forecasts from M_1 will be biased by the bias in the random walk forecast of $x_{2,T+1}$, so (see Appendix A.5 for derivations) neglecting the small impact of $\eta_{i,T}$ on $\beta_i - \widehat{\beta}_i$:

$$\mathrm{E}\left[\overline{\overline{\epsilon}}_{T+1|T}\right] = \beta_2 \delta,$$

and the resulting mean square forecast error is:

$$\mathrm{MSFE}_1 = \mathrm{E}\left[\overline{\overline{\epsilon}}^2_{T+1|T}\right] = \beta_2^2\delta^2 + 2\left(\beta_1^2 + \beta_2^2\right) + 4\rho\beta_1\beta_2 + \sigma_\epsilon^2\left(1 + 2T^{-1}\right). \tag{26}$$

Comparison with (23) highlights the additional cost of using the random walk relative to the in-sample mean when neither forecasting device can predict the break, since:

$$\mathrm{E}[\widehat{\widehat{\epsilon}}^2_{T+1|T}] - \mathrm{E}\left[\overline{\overline{\epsilon}}^2_{T+1|T}\right] = -\left(\beta_1^2 + \beta_2^2 + 2\rho\beta_1\beta_2 + 2\sigma_\epsilon^2 T^{-1}\right).$$

The in-sample mean of x_1 is the optimal forecast of $x_{1,T+1}$ given its in-sample stationarity, so irrespective of the value of β_2, the in-sample mean forecasts dominate when the shift is during the forecast period. When $\beta_2 = 0$, (26) collapses to $\approx \sigma_\epsilon^2 + 2\beta_1^2$, ignoring $\mathrm{O}_p(T^{-1})$ terms, compared to $\sigma_\epsilon^2 + \beta_1^2$ for the in-sample mean forecasts. A random walk doubles the error variance, so can be costly if there are no breaks or if the break occurs after the forecast origin. As for the in-sample mean case, the MSFE of M_1 is a function of the break.

The forecast bias for M_2 is the same as that for M_1 by the same argument, although MSFE_2 (reported in Appendix A.5) does deviate from that for M_1 as ψ_β^2 increases. This is due to the correlation parameter ρ which is picking up part of the omitted variable $x_{2,T+1}$ in M_2 and has more effect as ψ_β^2 increases. When $\beta_2 = 0$, $\text{MSFE}_2 \approx \sigma_\epsilon^2 + 2\beta_1^2$, which is the same as for M_1. Despite small but increasing deviations as ψ_β^2 increases, MSFE_2 follows a similar trajectory to MSFE_1. The misspecification is less relevant for the random walk forecasts of the marginal processes relative to the effect of the break, similar to the results for the in-sample mean forecasts.

5.6. Selecting Forecasted Regressors

In practice, selection will be applied to determine whether to include $x_{2,t}$ or not. Then, from (12), we can obtain the MSFE_3 as:

$$\text{MSFE}_3 = \text{MSFE}_1 + (1 - \text{p}_\alpha(\psi_\beta)) \left(\sigma_\epsilon^2 T^{-1} \left[\psi_\beta^2 \left\{ \frac{(1+\rho^2)}{(1-\rho^2)} + T^{-1} \right\} + 1 \right] \right).$$

The trade-off between parameter estimation uncertainty and including x_2 is essentially the same as in the known variable case: if x_2 has a noncentrality of zero, so $\beta_2 = \psi_\beta^2 = 0$, then the one-step MSFE is minimized by excluding x_2 from the forecasting model. It should be included if $\psi_\beta^2 > 1$. However, depending on the values of ρ and T, the switch point can be smaller than $\psi_\beta^2 = 1$, although the impact is likely to be small given the scale factor $\sigma_\epsilon^2 T^{-1}$. Even though the random walk forecast is highly uncertain by using just one observation, if the variable that breaks is quite significant then it pays to include that variable when using the random walk forecast.

Figure 6 also records the MSFEs for the random walk forecasts using the same parameter values. The increase in MSFE over the in-sample mean forecasts is evident. Both MSFE_1 and MSFE_2 follow similar trajectories, although they do start to diverge for large ψ_β^2, with MSFE_3 at $\alpha = 0.16$ close to MSFE_1.

6. An In-Sample Shift in the Regressors

In contrast to the previous section, the break is assumed to occur at T, which is the last observation available for estimation. Now there is information available regarding the break when the forecasts are made. Such a framework would also be relevant in sequential forecasting. We consider forecasting using in-sample means. In common with the previous section, we study selection (Sections 6.3 and 6.5), the random walk device to forecast the regressors (Section 6.4), and finally using the random walk to forecast y (Section 6.6).

6.1. Specification of the In-Sample Shift

The DGP is adapted from (15) but the shift in μ_2 occurs at T, rather than $T+1$:

$$\begin{aligned} x_{1,t} &= \mu_1 + \eta_{1,t} && t = 1, \ldots, T+1, \\ x_{2,t} &= \begin{cases} \mu_2 + \eta_{2,t} & t = 1, \ldots, T-1, \\ \mu_2 + \delta + \eta_{2,t} & t = T, T+1. \end{cases} \end{aligned} \tag{27}$$

6.2. Forecasting Regressors Using In-Sample Means

The relationship of interest, i.e., the conditional equation for y_{T+1}, remains constant. However, the in-sample mean μ_y is shifted to $(\mu_y + \beta_2\delta)$ at T. Although the only DGP parameter to shift is μ_2 to $\mu_2 + \delta$, sample calculations will be altered as now $\text{E}[\bar{x}_2] = \mu_2 + T^{-1}\delta$ (see Appendix A.6 for derivations).

The impact on the estimated in-sample mean of $\{x_{2,t}\}$ will be small from the break, unless δ is very large, so by using the in-sample means for their future unknown values,

the forecasted mean of y_{T+1} for M_1 will still be close to μ_y, and the resulting forecast error bias is:

$$\mathrm{E}\left[\widehat{\widehat{\epsilon}}_{T+1|T+1}\right] \approx \beta_2\delta\left(1-T^{-1}\right).$$

This is unbiased when $\beta_2 = 0$, but could be badly biased if $\beta_2\delta$ is large. The MSFE for M_1 is:

$$\mathrm{MSFE}_1 = \mathrm{E}\left[\widehat{\widehat{\epsilon}}^2_{T+1|T+1}\right] = \beta_2^2\delta^2\left(1-T^{-1}\right)^2 + \beta_1^2 + \beta_2^2 + \sigma_\epsilon^2. \tag{28}$$

This is very similar to the MSFE_1 in (23) for an out-of-sample break using the in-sample means to forecast the exogenous regressors, and hence MSFE_2 and MSFE_3 as well, although the correlation between the two regressors does not enter.

When $\beta_2 = 0$, both (23) and (28) collapse to $\sigma_\epsilon^2 + \beta_1^2$. The dampening of the squared location shift by $\left(1-T^{-1}\right)^2$ slightly improves the MSFE for the in-sample shift relative to an out-of-sample shift at larger ψ_β^2, as shown in Figure 7.

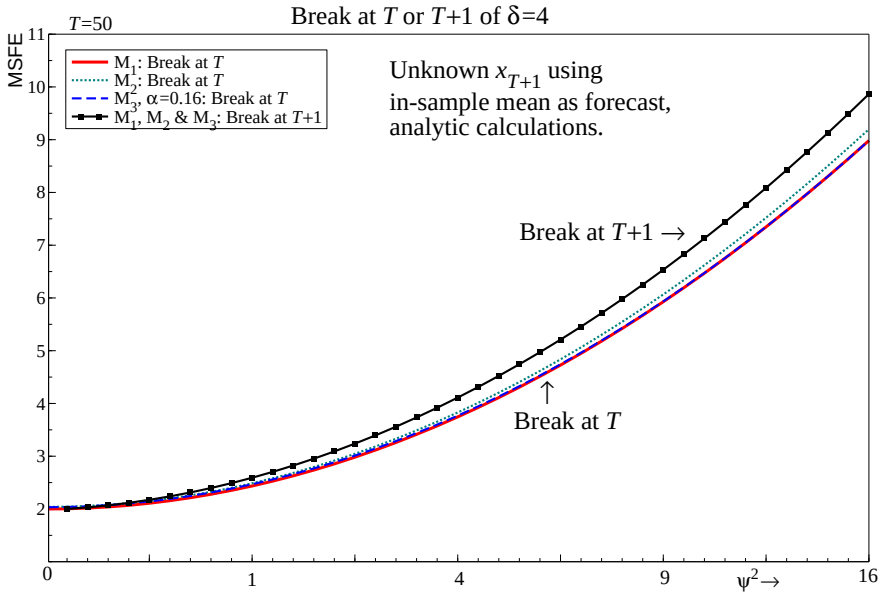

Figure 7. MSFE_1, MSFE_2, and MSFE_3 for unknown future exogenous regressors where the break occurs in the mean of x_2 at T and the in-sample mean is used as the forecast for the regressors. Included are the results when the break occurs at $T+1$.

For a break out of sample, we find the analytic results for M_2 are identical to those for M_1 (see Section 5.4). For the in-sample break, the forecast error and MSFE for M_2 does differ to that of M_1 (see Appendix A.6 for analytic results). This is because the in-sample location shift affects ρ which introduces a term similar to the squared location shift scaled by T in (28). Therefore, $\mathrm{MSFE}_1 \neq \mathrm{MSFE}_2$ unless $\beta_2 = 0$, with M_2 incurring a larger MSFE cost as ψ_β^2 increases due to misspecification, although the divergence is small even for small T, and disappears asymptotically.

6.3. Selecting Regressors

Selection follows from (12) and hence:

$$\mathrm{MSFE}_3 \approx \mathrm{MSFE}_1 + \left(1-\mathrm{p}_\alpha\left(\psi_\beta\right)\right)\left[\sigma_\epsilon^2 - \beta_1^2 - \rho^2\beta_2^2 + 2T^{-1}\left(\sigma_v^2 + \beta_2^2\delta^2\right)\right].$$

The cost of omitting x_2 rises with $\beta_2^2\delta^2$, although increases in β_2 will raise ψ_β^2 and hence raise the probability of retaining x_2, albeit unconnected with the magnitude of δ^2. As the location shift is scaled by T, $\mathrm{MSFE}_3 \to \mathrm{MSFE}_1$ as $T \to \infty$.

6.4. Forecasting Regressors Using a Random Walk

From the previous analysis in Section 6.2, knowledge of the break at T brought little benefit when using in-sample means as forecasts. However, the random walk should do better when the break occurs at T as opposed to $T+1$. As before:

$$\widetilde{\widetilde{x}}_{1,T+1|T} = x_{1,T} \quad \text{and } \widetilde{\widetilde{x}}_{2,T+1|T} = x_{2,T},$$

but now $\mathrm{E}[x_{1,T}] = \mu_1$ and $\mathrm{E}[x_{2,T}] = \mu_2 + \delta$, and hence $\mathrm{E}[\Delta x_{1,T+1}] = 0$ and $\mathrm{E}[\Delta x_{2,T+1}] = 0$ as well.

Given the unbiased forecasts of the exogenous regressors, it follows that the forecasts for M_1 are unbiased (see Appendix A.7) when the parameter estimates are unbiased. The MSFE for M_1 is:

$$\mathrm{MSFE}_1 = \mathrm{E}\left[\widetilde{\widetilde{\epsilon}}^2_{T+1|T}\right] = 2\left(\beta_1^2 + \beta_2^2\right) + 4\rho\beta_1\beta_2 + \sigma_\epsilon^2\left(1 + \frac{2}{T} + \frac{\delta^2}{T(1-\rho^2)}\right). \tag{29}$$

When $\beta_2 = 0$, the MSFE is similar to that of the out-of-sample break case, where the random walk is costly as forecasts of both $x_{1,T+1}$ and $x_{2,T+1}$ are inefficient. However, (29) does depend on the magnitude of the shift independently of β_2, unlike (26). MSFE_1 is a function of ψ_β^2, increasing as ψ_β^2 increases, unlike in the known regressor case. But it does so more slowly than for breaks out of sample, or breaks in sample using the in-sample mean. As ψ_β^2 increases, the break at T in μ_2 has a larger effect on the dependent variable, and hence the benefits of using a random walk forecast of $x_{2,T+1}$ are larger.

M_2 will suffer when $\beta_2 \neq 0$ as the forecasts will be biased. The MSFE for M_2 is:

$$\mathrm{MSFE}_2 = \mathrm{E}\left[\widetilde{\widetilde{\epsilon}}^2_{T+1|T}\right] = \beta_2^2\left(\delta^2 + \rho^2 + 1\right) + 2\beta_1^2 + 4\rho\beta_1\beta_2 + \sigma_\epsilon^2\left(1 + T^{-1} + T^{-2}\psi_\beta^2\right), \tag{30}$$

so no robustness in the sense of reducing bias is achieved unless $\beta_2 = 0$. When $\beta_2 = 0$, $\mathrm{MSFE}_2 < \mathrm{MSFE}_1$, but the bias from not including a random walk, and hence unbiased, forecast of $x_{2,T+1}$ quickly outweighs parameter estimation costs as ψ_β^2 increases.

Solving for $\mathrm{MSFE}_2 < \mathrm{MSFE}_1$ results in:

$$\psi_\beta^2 < \frac{(1-\rho^2) + \delta^2}{(1-\rho^2)(T^{-1}-1) + \delta^2}. \tag{31}$$

The break term dominates and offsets on the numerator and denominator, leading to a trade-off at $\approx$1 with deviations scaled by T^{-1}. For $\rho = 0.5$, $T = 100$ and $\delta = 4$, MSFE_2 dominates when $\psi_\beta = 1.05$. Interestingly, the cut-off is slightly above 1 for this case, compared to slightly below 1 for the known breaks out-of-sample case, but the results still imply that a selection significance level of approximately 16% would be optimal to trade-off the cost of estimating an additional parameter.

Figure 8 records the MSFEs from M_1 (29), M_2 (30) and three values of M_3 (A4) for the analytic results. There is a clear trade-off at $\psi_\beta^2 \approx 1$, just as in the known breaks case.

6.5. Selecting Forecasted Regressors

The final step is to compute the MSFE for M_3 for the random walk forecast, reported in Appendix A.7. Just as regression models are usually selected, that will occur for any forecasting devices designed to minimize systematic bias. As with Figure 5, selection between M_1 and M_2 can be advantageous even for these forecasting devices as seen in Figure 8. Selection outperforms M_1 for $\psi_\beta^2 < 1$, and remains close to the MSFE_1 at $\alpha = 0.05$ and $\alpha = 0.16$, again in all cases matching or outperforming always using M_2.

A comparison with the MSFE for the in-sample mean forecasts, also recorded in Figure 8, suggests a possible forecast improvement. If the regressor that breaks at T is known, combining the in-sample mean forecast for M_1 with the random walk forecast

for M_2 will improve forecast performance (shifting the MSFE curves for the random walk forecast down by approximately 1). As the number of regressors increases, the forecasting method for each contemporaneous regressor will have a cumulative impact. However, as the break occurs in sample, methods to detect breaks at the forecast origin such as impulse indicator saturation (IIS) could be used to guide the forecaster to the most appropriate forecasting device.[4] Selection between forecasting devices that minimize systematic bias versus those that trade-off bias and variance requires pre-testing and would only help for in-sample shifts; see, e.g., Chu et al. (1996).

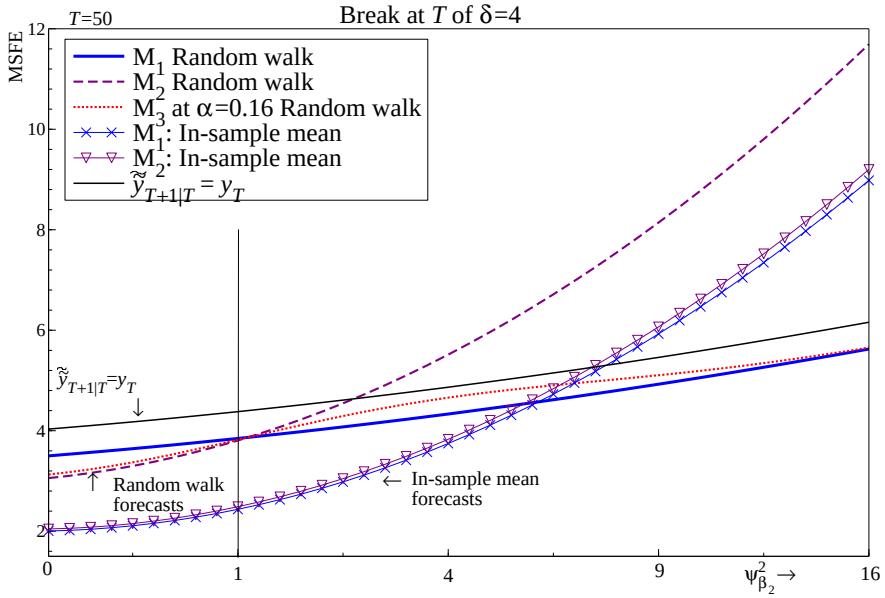

Figure 8. MSFE comparisons between M_1, M_2 and M_3 at $\alpha = 0.16$ for unknown future exogenous regressors where the break occurs in the mean of x_2 at T and the last in-sample observation is used as the forecast for the conditioning regressors. Also recorded is the MSFE for M_1 and M_2 using in-sample means and a misspecified random walk for y_{T+1} directly.

Thus, selection can be valuable for forecasting to the extent that it retains relevant regressors that shift (here, x_2), and also if it eliminates irrelevant regressors that shift, as considered in Section 9.

6.6. Forecasting the Dependent Variable Using a Random Walk

If a break is suspected, an alternative to the approaches considered so far is to use a knowingly misspecified model of the conditional DGP. One possibility is to use a random walk forecast for y, with the advantage that y_T is known and avoids the need to forecast $x_{1,T+1}$ and $x_{2,T+1}$. Hendry and Mizon (2012) derive a forecast-error taxonomy for open models that demonstrates the numerous additional forecast errors that arise from forecasting regressors offline in open models. They show that, in some cases, it can pay to use a misspecified model rather than to forecast the regressors offline. The forecast device is:

$$\widetilde{\widetilde{y}}_{T+1|T} = y_T.$$

Then

$$y_T = \mu_y + \beta_2\delta + \beta_1\eta_{1,T} + \beta_2\eta_{2,T} + \epsilon_T$$

is a noisy one-observation estimator of $(\mu_y + \beta_2\delta)$. The outturn at $T+1$ is:

$$y_{T+1} = (\mu_y + \beta_2\delta) + \beta_1\Delta\eta_{1,T+1} + \beta_2\Delta\eta_{2,T+1} + \epsilon_{T+1} + \beta_1\eta_{1,T} + \beta_2\eta_{2,T}.$$

The forecast error is given by:

$$\widetilde{\widetilde{e}}_{T+1|T} = y_{T+1} - \widetilde{\widetilde{y}}_{T+1|T} = \beta_1 \Delta\eta_{1,T+1} + \beta_2 \Delta\eta_{2,T+1} + \Delta\epsilon_{T+1},$$

which is unbiased and has a MSFE of:

$$\text{MSFE}_4 = \mathrm{E}\left[\widetilde{\widetilde{e}}^2_{T+1|T}\right] = 2\left(\beta_1^2 + \beta_2^2\right) + 4\rho\beta_1\beta_2 + 2\sigma_\epsilon^2.$$

This is independent of δ so should perform relatively the best when δ^2 is large, although performs worse than random walk forecasts for $x_{1,T+1}$ and $x_{2,T+1}$ when ψ_β^2 is small; see Figure 8. The forecasts are invariant to omitting x_2 since this random walk forecast is independent of the regressors, which is a major advantage and negates the role of selection. However, there is a cost when the model is correctly specified. The results in the simulation below suggest that such an approach should be viewed as complementary, with forecast pooling across selected conditional models and misspecified robust devices designed to mitigate bias frequently outperforming individual methods.

7. Summary of Analytic Results and the Impact of Selection

The theoretical analysis has established four results.

1. Regressors should be retained if $\psi_\beta \geq 1$. This is established for DGPs that are stationary or with a break out of sample for known regressors and a break in sample for random walk forecasts.
2. For the two-regressor case, $\psi_\beta = 1$ maps to $\alpha \approx 0.16$. Selection delivers improvements to the one-step-ahead MSFE for $\psi_\beta < 1$ and can be close to the correct model specification for $\psi_\beta > 1$, with the largest deviations occurring at intermediate values of ψ_β.
3. If there are breaks out of sample and contemporaneous regressors need to be forecast, the break dominates the MSFE and selection plays almost no role. Similar results are found even if the break occurs at the end of the sample, but the in-sample mean is used to forecast the regressors.
4. Random walk forecasts are costly if there are no breaks (forecasting $x_{1,T+1}$) or if the breaks are unpredictable (a break at $T+1$ and forecasting $T+1|T$). However, they improve MSFE when the break is predictable (break at T and forecasting $T+1|T$).

Table 3 summarises the results for specific parameters using $T = 50$ ($T = 100$ is in Table A1 in Appendix B). For each scenario, the ratio of $\text{MSFE}_j/\text{MSFE}_1$ for $j = 2, 3$ is reported. MSFE_2 has no selection, and is therefore listed as $\alpha = 0$, while three values of α are used for MSFE_3. The squared noncentralities $\psi_\beta^2 = 0, 1, 4, 9, 16$ capture the full hump shape seen in the figures above.

M_2 is the correct model in the column labelled $\psi_\beta^2 = 0$, so the ratio of $\text{MSFE}_2/\text{MSFE}_1$ measures the cost of over-specification. The gains can be substantial in some cases, almost 30% for a break out of sample with known regressors, but in other cases including $x_{2,t}$ is not at all costly despite its irrelevance. Tighter selection for M_3 is close to M_2 as $x_{2,t}$ will be omitted more frequently, but even at $\alpha = 0.16$ the ratio for M_3 is close to the ratio for M_2, suggesting that selection is not costly.

Moving to the next column highlights the $\psi_\beta = 1$ trade-off, with all cases almost exactly equal to one. A cut-off slightly lower than one was found in (19), which is reflected in the ratio marginally greater than one. Conversely, (31) found a cut-off slightly larger than one, resulting in a ratio slightly below one, but the differences are small.

Next, consider the columns labelled $\psi_\beta^2 = 4, 9$, and 16. M_1 is the correct model so the objective is to minimize the ratio. In some cases M_2 performs poorly, but M_3 at $\alpha = 0.16$ is frequently very close to 1, i.e., MSFE_1. Selection forecast performance tends to be worse at $\psi_\beta^2 = 4$, but as the signal for x_2 increases, the probability of retaining x_2 increases so the selected model is closer to M_1. The benefits of selection vary by case. For example, for a break at T using in-sample means, selection at $\alpha = 0.16$ delivers a 2.4% improvement

relative to M_2 for $\psi_\beta = 4$, compared to a halving of the ratio for the random walk. In almost every setting, $MSFE_3$ is close to $MSFE_1$ so the costs of selection are usually small, irrespective of the noncentrality. In that sense, model selection acts to reduce the risk relative to the worst model. Conversely, the costs of unmodeled shifts are very large, up to almost 8-fold greater than the baseline stationary $MSFE_1$.

Table 3. Ratio of MSFE to that of $MSFE_1$, $T = 50$. M_2 has no selection ($\alpha = 0$); selection in M_3 at α.

	MSFE Relative to $MSFE_1$				
Model	$\psi_\beta^2 = 0$	$\psi_\beta^2 = 1$	$\psi_\beta^2 = 4$	$\psi_\beta^2 = 9$	$\psi_\beta^2 = 16$
Sections 4.1 and 4.2 No shift with known future regressors					
$\alpha = 0$ (M_2)	0.981	1.001	1.060	1.158	1.295
$\alpha = 0.001$	0.981	1.000	1.051	1.093	1.068
$\alpha = 0.05$	0.982	1.000	1.027	1.023	1.006
$\alpha = 0.16$	0.984	1.000	1.016	1.008	1.001
Sections 5.2 and 5.3 Out-of-sample shift with known future regressors					
$\alpha = 0$ (M_2)	0.709	1.014	1.927	3.450	5.582
$\alpha = 0.001$	0.709	1.013	1.836	2.505	2.095
$\alpha = 0.05$	0.724	1.011	1.449	1.366	1.095
$\alpha = 0.16$	0.756	1.009	1.256	1.136	1.022
Section 5.4 Out-of-sample shift with mean forecast of future regressors					
$\alpha = 0$ (M_2)	1.000	1.000	1.000	1.000	1.000
$\alpha = 0.001$	1.000	1.000	1.000	1.000	1.000
$\alpha = 0.05$	1.000	1.000	1.000	1.000	1.000
$\alpha = 0.16$	1.000	1.000	1.000	1.000	1.000
Section 5.5 Out-of-sample shift with random walk forecast of future regressors					
$\alpha = 0$ (M_2)	0.993	1.004	1.020	1.034	1.043
$\alpha = 0.001$	0.993	1.004	1.018	1.021	1.010
$\alpha = 0.05$	0.994	1.003	1.010	1.005	1.001
$\alpha = 0.16$	0.994	1.002	1.006	1.002	1.000
Sections 6.2 and 6.3 In-sample shift with mean forecast of future regressors					
$\alpha = 0$ (M_2)	1.020	1.021	1.022	1.023	1.024
$\alpha = 0.001$	1.020	1.021	1.020	1.014	1.006
$\alpha = 0.05$	1.019	1.017	1.011	1.004	1.000
$\alpha = 0.16$	1.017	1.014	1.006	1.001	1.000
Sections 6.4 and 6.5 In-sample shift with random walk forecast of future regressors					
$\alpha = 0$ (M_2)	0.871	0.990	1.273	1.653	2.078
$\alpha = 0.001$	0.871	0.990	1.246	1.401	1.258
$\alpha = 0.05$	0.878	0.991	1.132	1.097	1.022
$\alpha = 0.16$	0.892	0.993	1.075	1.036	1.005

These results show that even facing breaks, the well-known trade-off for selecting variables in forecasting models, namely that variables should be retained if their noncentralities exceed 1, still applies, resulting in much looser significance levels than typically used. The problem with such an approach is that when many $\beta_{2,i} = 0$ but are subject to location shifts, M_1, which erroneously includes $x_{2,t}$ in the model, will perform worse. Loose significance levels increase the chance that irrelevant variables with $\psi_\beta = 0$ are retained by being adventitiously significant for that draw. To evaluate this effect, the next section undertakes a simulation study of selection in models with ten irrelevant and five relevant exogenous regressor variables confronting a variety of shifts.

8. Simulation Design

We generalize the above analysis using Monte Carlo analysis, formalizing the DGP and models that are estimated. We consider larger models with dynamics, evaluating for a

range of strategies to forecast future values of the regressors, different significance levels, and different configurations of out-of-sample breaks. The next section then evaluates the simulation results.

8.1. Data Generation Process

The DGP is for a scalar dependent variable y_t, and N regressors $\mathbf{x}_t = (x_{1,t}, \ldots, x_{N,t})'$. There are n regressors that are relevant, i.e., have a nonzero coefficient in the DGP for y_t, and $N - n$ that are irrelevant with coefficient zero.

We wish to introduce breaks either in relevant, or irrelevant, or both types of regressors. For convenience we assume that the regressors are ordered by increasing significance (i.e., squared noncentrality $\psi^2_{\beta_i}$). The DGP for y is an AR(1) with regressors:

$$y^*_t = \beta_0 + \beta_y y^*_{t-1} + \sum_{j=1}^{N} \beta_j x_{j,t} + \epsilon_t, \quad \epsilon_t \sim \mathrm{IN}[0, \sigma^2_\epsilon], \quad t = -Q+1, \ldots, 0, 1, \ldots, T+H. \tag{32}$$

The regressors are independent of each other and (in sample) have a common autoregressive coefficient λ and mean $\delta/(1-\lambda)$. We allow for a break in observations $T+1$ and $T+2$, using subscript I if the break applies to xs that are irrelevant in (32) (i.e., have a coefficient of zero) and R for those that are relevant:

$$\begin{aligned}
x_{j,t} &= \delta + \lambda x_{j,t-1} + \eta_{j,t}, \quad \eta_{j,t} \sim \mathrm{IN}[0,1], && j = 1, \ldots, N, \; t = -Q+1, \ldots, T, T+3, \ldots, \\
x_{j,t} &= \delta_I + \lambda_I x_{j,t-1} + \eta_{j,t}, \quad \eta_{j,t} \sim \mathrm{IN}[0,1], && j = 1, \ldots, N-n, \; t = T+1, T+2, \\
x_{j,t} &= \delta_R + \lambda_R x_{j,t-1} + \eta_{j,t}, \quad \eta_{j,t} \sim \mathrm{IN}[0,1], && j = N-n+1, \ldots, \; t = T+1, T+2.
\end{aligned} \tag{33}$$

Throughout, we set $\sigma^2_\epsilon = 1$, $\beta_0 = 5$, $\beta_y = 0.5$, $\delta = 2$, $N = 15$. Fifty initial observations are discarded ($Q = 50$). We set observation zero equal to twenty in each replication, giving the generated data as:

$$y_t = y^*_t + 20 - y^*_0. \tag{34}$$

The remaining coefficients in (32) are specified through their noncentralities. We run three alternative experiments:

$$\begin{aligned}
\boldsymbol{\psi}(1): &\quad \boldsymbol{\psi}_\beta = (0,0,0,0,0,1.2,1.2,1.2,1.2,1.2,1.2,1.2,1.2,1.2,1.2)', \\
\boldsymbol{\psi}(2): &\quad \boldsymbol{\psi}_\beta = (0,0,0,0,0,0,0,0.5,1,1.5,2,3,4)', \\
\boldsymbol{\psi}(4): &\quad \boldsymbol{\psi}_\beta = (0,0,0,0,0,0,0,0,0,0,0,0,4,4,4)'.
\end{aligned} \tag{35}$$

Then $\beta_j = \psi_{\beta_j} \{(T-N-2)\mathrm{V}[x_{j,t}]\}^{-1/2}$, using the in-sample variances computed over $t = 1, \ldots, T$. This ensures that the t-values in the estimates of (32) will be equal to ψ_{β_j} on average. Note that the noncentralities in each specification sum to twelve, and have $n = 10, 6, 3$ respectively.

With common coefficients δ and λ, the regressors are exchangeable in analytical calculations. The unconditional process of each x_j, in the absence of any break, has mean $\overline{x} = \delta/(1-\lambda)$ and variance $(1-\lambda^2)^{-1}$. When $\delta = 2$ and $\lambda = 0.75$, the steady state for y^*_t is then $\overline{y} = 10 + 2 \times 8 \times 12 \times \{83/(1-0.75^2)\}^{-1/2} \approx 10 + 16 \times 0.87 = 23.9$, using total noncentrality of 12, $\overline{x} = 8$, $T = 100$, $N = 15$. The degrees-of-freedom adjustment counts N, the intercept, and the lagged dependent variable.

Breaks in the process for the target variable y are introduced through breaks in the regressors. During the break, $\delta_R = -0.3 \equiv \delta_\Delta$, so δ drops by -2.3. Keeping λ unchanged, the equilibrium changes from $\overline{x} = 8$ to $\overline{x}_\Delta = -1.2$, which is a shock of six unconditional standard errors. The impact on y_t depends on the coefficients β_j. To quantify this, it is convenient to assume that the processes are at their unconditional means, after which we follow the shocks through the dynamic system, ignoring the disturbances. The impact on x when the coefficients change from $(\delta, \lambda) = (2, 0.75)$ to $(\delta_\Delta, \lambda_\Delta)$ is given in Table 4.

Table 4. Impact on x when coefficients change from $(\delta, \lambda) = (2, 0.75)$ to $(\delta_\Delta, \lambda_\Delta)$.

$(\delta_\Delta, \lambda_\Delta) =$	**(2, 0.75)**	**(−0.3, 0.75)**	**(−0.3, 0.95)**	**(−0.3, 0.05)**	**(2, 0.05)**	**(2, 0.95)**
$x_{j,T+1\|T}$	8	5.7	7.3	0.1	2.4	9.6
$x_{j,T+2\|T+1}$	8	4.0	6.6	−0.3	2.1	11.1
$x_{j,T+3\|T+2}$	8	5.0	7.0	1.8	3.6	10.3

The process reverts to the original coefficients at $T + 3$, aiming to capture qualitatively aspects of a sustained but temporary structural break, such as the Great Financial Crisis or the COVID-19 pandemic. The impact of the break on $y_{j,T+1|T}$ is 0.87 times the new x. For $(-0.3, 0.95)$ this is a change of 0.6, well below y's conditional standard error of unity.

Table 5 lists the break settings we consider. The upward break in slope (a) pushes the process towards a unit root, while the downward break in slope (b) makes it almost white noise. Figure 9 plots the second half of y_t for one replication of the DGP and for each of the five specifications of the break. This is for $T = 100$ and after discarding the initial observations. The break lasts for two observations in the forecast period, after which the DGP reverts to the settings without break. Figure 9 illustrates the low impact of the break in mean and slope when $(\delta_\Delta, \lambda_\Delta) = (-0.3, 0.95)$.

Table 5. Configurations of breaks in the simulations.

	δ_Δ	λ_Δ
No break	2	0.75
Break in mean	−0.3	0.75
Break in slope (a)	2	0.95
Break in slope (b)	2	0.05
Break in mean and slope (a)	−0.3	0.95
Break in mean and slope (b)	−0.3	0.05

The design (33) allows for breaks in relevant variables, in irrelevant variables, or in both. In the last case: $\delta_R = \delta_I = \delta_\Delta$ and $\lambda_R = \lambda_I = \lambda_\Delta$. Breaks in irrelevant variables do not affect y, but can have an impact on forecasts if the irrelevant variables are used in the forecasts' construction. However, when forecasting for $T + 1|T$, such breaks have no impact at all, because the future x_{T+1}s are not yet known.

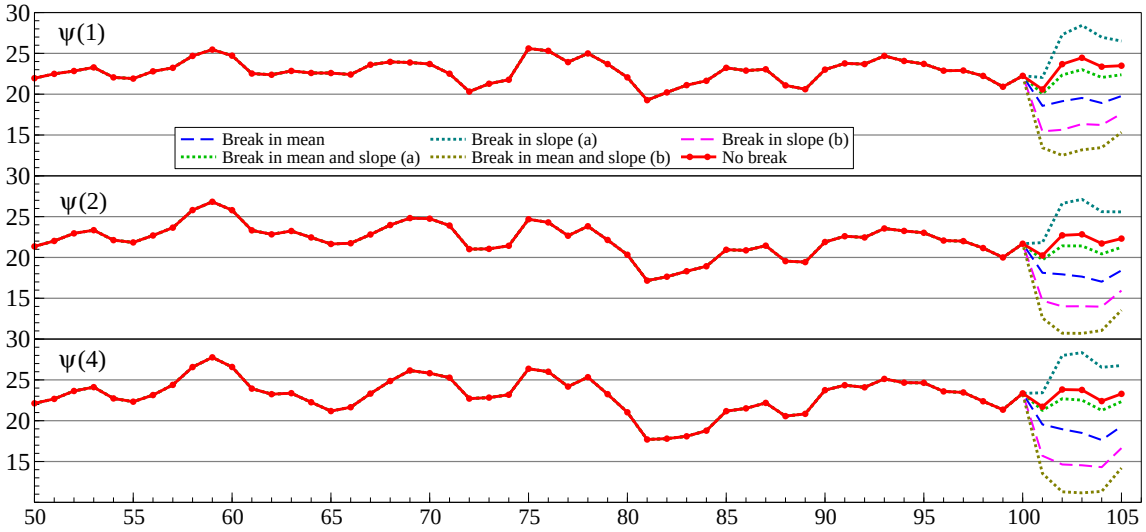

Figure 9. One replication of the DGP without break (solid line) and breaks as in Table 5, $T = 100$, $H = 5$.

8.2. Models and Forecast Devices

We generate $Q + T + H$ observations from DGP (32)–(34), discarding the initial Q. The starting point for modeling is the general unrestricted model (GUM):

$$y_t = \beta_0 + \beta_y y_{t-1} + \sum_{j=1}^{N} \beta_j^* x_{j,t} + \sum_{j=1}^{N} \gamma_j^* x_{j,t-1} + \epsilon_t, \quad \text{for } t = 1, \ldots, T. \tag{36}$$

An asterisk indicates that model selection is used, so the intercept and lagged y are not selected over but are always retained. Model selection is only performed once for each replication, but the selected model is re-estimated by ordinary least squares (OLS) each time that we forecast given data up to $T+h-1$:

$$y_t = \beta_0 + \beta_y y_{t-1} + \sum_{\widehat{\beta}_j^* \neq 0} \beta_j x_{j,t} + \sum_{\widehat{\gamma}_j^* \neq 0} \gamma_j x_{j,t-1} + \varepsilon_t, \quad \text{for } t = h, \ldots, T + h - 1. \tag{37}$$

Only one-step-ahead forecasts are generated and evaluated:

$$\widehat{y}_{T+h|T+h-1} = \widehat{\beta}_0 + \widehat{\beta}_y y_{T+h-1} + \sum_{\widehat{\beta}_j^* \neq 0} \widehat{\beta}_j \widetilde{x}_{j,T+h} + \sum_{\widehat{\gamma}_j^* \neq 0} \widehat{\gamma}_j x_{j,T+h-1} \text{ for } h = 1, ..., H. \tag{38}$$

The out-of-sample values $\widetilde{x}_{j,T+h}$ of the regressors in (38) are unknown when forming the forecasts. We consider a range of forecast devices that can supply these missing values:

INF: future outcomes: $\widetilde{x}_{j,T+h} = x_{j,T+h}$;

AVG: the in-sample average: $\widetilde{x}_{j,T+h} = \sum_{t=h}^{T+h-1} x_{j,t}/T$;

ARX: an AR(1) for each regressor: $\widetilde{x}_{j,T+h} = \widehat{\mu}_j + \widehat{\rho}_j x_{j,T+h-1}$, estimated by OLS for each horizon from:

$$x_{j,t} = \mu_j + \rho_j x_{j,t-1} + u_{j,t}, \quad t = h, \ldots, T + h - 1; \tag{39}$$

RWX: the random walk forecast: $\widetilde{x}_{j,T+h} = x_{j,T+h-1}$;

RDX: a random walk with differencing (Hendry 2006), using differenced estimates from (39):

$$\widetilde{x}_{j,T+h} = x_{j,T+h-1} + \widehat{\rho}_j \Delta x_{j,T+h-1}.$$

CAX: Cardt forecast of $\widetilde{x}_{j,T+h}$.

In addition, several alternatives that ignore the regressors are considered:

RWY: a random walk forecast: $\widehat{y}_{T+h} = y_{T+h-1}$;

ARY: an AR(1) forecast: $\widehat{y}_{T+h} = \widehat{\gamma}_0 + \widehat{\gamma}_1 y_{T+h-1}$, estimated by OLS for each horizon;

CAY: Cardt forecasts of $\widehat{y}_{T+h}$.

Model selection is performed using Autometrics (Doornik 2009) for a range of target significance levels $\alpha = (0.001, 0.01, 0.05, 0.1, 0.16, 0.32)$. Forecasting from a re-estimated GUM (37) without selection is also considered (i.e., $\alpha = 1$). Dropping all regressors (i.e., $\alpha = 0$) leaves the AR(1) model for y_t.

The devices that forecast the regressors supply plug-in values to allow forecasting with the GUM (36), as well as the reductions (37) of the GUM, at a range of nominal significance levels. Device INF uses future outcomes, making it infeasible for stochastic variables. Note that all devices using regressors benefit from some knowledge that is not available in practice, namely that the DGP is nested in the GUM, and the GUM is not misspecified. The fact that the regressors are exchangeable and break at the same time in the same way may also help: finding just one that matters could already improve the forecasts.

Cardt is a slightly improved version of Card (calibrated average of rho and delta methods), see Doornik et al. (2020a), which performed very well in the M4 forecast competition of Makridakis et al. (2020). Cardt averages forecasts from a differenced, autoregressive, and a moving average model. These are then treated as future observations in a calibration model with richer autoregressive structure. The full procedure is documented in Castle et al. (2021). Cardt pays particular attention to seasonality, which is irrelevant here. We use Cardt to make four forecasts, then use the first of these. The method will take logarithms by default. Switching that off makes little difference in these experiments. Cardt is used in daily COVID-19 forecasts of Doornik et al. (2020b).

8.3. Selecting Regressors

The noncentrality ψ_β in the DGP affects the probabilities of retaining a variable in the model selection procedure. Table 6 shows the probability of retaining one or all relevant regressors assuming independent t-tests. While the probability of retaining one variable may be quite large, the joint probability of retaining all can be extremely low. Thus, even using a significance level of 16%, many relevant variables will be omitted if their noncentralities are small. However, their contribution to explaining the dependent variable is also small and breaks in such variables will have a smaller effect.

Table 6. Probability of retaining one or all variables when the coefficients have the specified noncentrality, assuming independence at nominal significance α and Student-t(83) distribution.

	$\psi_\beta = 1.2$		$\psi_\beta = 0.5$	$\psi_\beta = 1$	$\psi_\beta = 1.5$	$\psi_\beta = 2$	$\psi_\beta = 3$	$\psi_\beta = 4$	Joint	Average	$\psi_\beta = 4$	
α	$n = 1$	$n = 10$	$n = 1$	$n = 1$	$n = 1$	$n = 1$	$n = 1$	$n = 1$	$n = 6$	$n = 6$	$n = 1$	$n = 3$
0.001	0.015	0.000	0.002	0.009	0.030	0.081	0.341	0.721	0.000	0.197	0.721	0.375
0.01	0.077	0.000	0.018	0.053	0.130	0.263	0.641	0.912	0.000	0.336	0.912	0.758
0.05	0.216	0.000	0.070	0.163	0.313	0.504	0.843	0.976	0.001	0.478	0.976	0.930
0.1	0.322	0.000	0.124	0.254	0.435	0.631	0.907	0.989	0.008	0.557	0.989	0.968
0.16	0.414	0.000	0.181	0.339	0.533	0.719	0.941	0.994	0.022	0.618	0.994	0.983
0.32	0.579	0.004	0.309	0.500	0.691	0.840	0.976	0.998	0.087	0.719	0.998	0.995

The fraction of relevant variables that is retained in the Monte Carlo experiment is denoted the potency, and the fraction of irrelevant variables that is retained is denoted the gauge. We always retain the intercept and lagged y, so the GUM (36) has $2N$ possible variables to select over, of which n are relevant. For $m = 1, \dots, M$ replications we define the indicator function $\mathbf{1}\{\cdot\}$ and:

$$\begin{aligned} \text{gauge}_m &= \frac{1}{2N-n}\left[\sum_{j=1}^{N-n} \mathbf{1}\{\widehat{\beta}_{j,m} \neq 0\} + \sum_{j=1}^{N} \mathbf{1}\{\widehat{\gamma}_{j,m} \neq 0\}\right], \\ \text{potency}_m &= \frac{1}{n}\sum_{j=N-n+1}^{N} \mathbf{1}\{\widehat{\beta}_{j,m} \neq 0\}. \end{aligned}$$

This is then averaged over all replications.

Table 7 shows that the empirical gauge matches the theoretical probabilities in Table 6 when using Autometrics for selection: the gauge is higher than α but not by much. Potencies are close to the powers of one-off t-tests with the same noncentralities, up to $\alpha = 0.1$, beyond that they fall behind. Consequently, it is appropriate to use Autometrics to investigate the theoretical results by simulating a more general setting, without concern that the selection algorithm will influence the results relative to the single t-test approach analyzed above.

Table 7. Gauge and potency for three noncentrality designs, M = 10,000 replications.

	Gauge			Potency		
α	$\psi(1)$	$\psi(2)$	$\psi(4)$	$\psi(1)$	$\psi(2)$	$\psi(4)$
0.001	0.005	0.006	0.006	0.034	0.205	0.712
0.01	0.025	0.024	0.020	0.113	0.345	0.884
0.05	0.079	0.075	0.069	0.231	0.458	0.919
0.1	0.126	0.124	0.121	0.297	0.507	0.919
0.16	0.181	0.180	0.178	0.355	0.545	0.923
0.32	0.328	0.328	0.327	0.479	0.634	0.941

9. Simulation Evidence

Simulation evidence is presented using the design of Section 8.1 and forecast devices of Section 8.2. All experiments use M = 10,000 and are implemented in Ox 9 (Doornik 2018) and PcGive (Hendry and Doornik 2018). We start with out-of-sample forecasts in Section 9.1, when the break is unanticipated. Then Section 9.2 compares breaks in relevant and irrelevant variables, Section 9.3 looks at forecasts after the break, Section 9.4 considers selection, Section 9.5 introduces pooled forecasts, and Section 9.6 summarizes.

9.1. Forecasting before the Break

The top half of Table 8 is for the case without breaks, when forecasting $T+1|T$ is similar to forecasting $T+2|T+1$, etc. The table reports the ratio of the MSFE for devices INF, AVG, ARX, RWX respectively to the MSFE of ARY for a range of significance levels α. Selection at $\alpha = 0$ implies dropping all the regressors, leaving an AR(1) in y, denoted ARY. The bottom row of each half gives the MSFE of ARY. Not selecting at all ($\alpha = 1$) coincides with the GUM.

Table 8. No break and out-of-sample break. Ratio of MSFE to MSFE_{ARY} forecasting $T+1|T$.

	$\psi(1)$				$\psi(2)$				$\psi(4)$			
	INF	**AVG**	**ARX**	**RWX**	**INF**	**AVG**	**ARX**	**RWX**	**INF**	**AVG**	**ARX**	**RWX**
Ratio	**No break**											
$\alpha = 0$	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
$\alpha = 0.001$	1.03	1.01	1.02	1.03	0.95	1.06	0.99	1.02	0.83	1.13	0.94	0.99
$\alpha = 0.01$	1.08	1.06	1.05	1.08	0.93	1.11	0.98	1.02	0.79	1.17	0.93	0.97
$\alpha = 0.05$	1.13	1.13	1.08	1.12	0.95	1.19	0.99	1.03	0.83	1.23	0.95	0.99
$\alpha = 0.1$	1.16	1.18	1.09	1.13	0.99	1.23	1.01	1.06	0.87	1.27	0.97	1.02
$\alpha = 0.16$	1.19	1.23	1.11	1.15	1.01	1.28	1.04	1.08	0.91	1.31	1.00	1.04
$\alpha = 0.32$	1.25	1.36	1.15	1.19	1.09	1.38	1.09	1.13	0.99	1.41	1.05	1.09
GUM	1.34	1.51	1.20	1.23	1.18	1.50	1.13	1.17	1.08	1.52	1.10	1.14
MSFEARY	1.15				1.31				1.43			
Ratio	**Average over five break types in relevant regressors**											
$\alpha = 0$	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
$\alpha = 0.001$	0.90	1.00	1.00	1.01	0.58	1.02	0.99	1.00	0.37	1.05	0.97	0.98
$\alpha = 0.01$	0.74	1.01	1.01	1.02	0.42	1.04	0.99	0.99	0.28	1.06	0.97	0.97
$\alpha = 0.05$	0.57	1.03	1.01	1.02	0.37	1.06	0.99	0.99	0.28	1.08	0.97	0.98
$\alpha = 0.1$	0.52	1.05	1.02	1.03	0.37	1.07	0.99	1.00	0.29	1.09	0.98	0.98
$\alpha = 0.16$	0.50	1.06	1.02	1.03	0.37	1.08	1.00	1.01	0.30	1.10	0.99	0.99
$\alpha = 0.32$	0.48	1.10	1.03	1.04	0.38	1.11	1.01	1.02	0.32	1.13	1.00	1.01
GUM	0.49	1.13	1.05	1.05	0.41	1.14	1.03	1.03	0.35	1.16	1.01	1.02
MSFEARY	18.58				18.80				18.98			

Without a break, knowing the future value of regressors, device INF, is only useful when they are significant. Using the sample mean AVG never improves one-step forecasting

relative to ARY. This also holds when there is a break, and is even more pronounced for $T+2|T+1$ and $T+3|T+2$ (not shown). We see that MSFE_{ARY} increases when there are more highly significant variables. There is an improvement over ARY from forecasting the regressors with ARX at strict significance levels for $\psi(4)$. In this stationary DGP without breaks, ARX dominates RWX: it is better to model the regressors by an autoregression (the true model) than taking the last known value.

The bottom half of Table 8 is for the cases with an out-of-sample break in the relevant variables only. The ratios for the five break settings (in mean, in slope, and in mean and slope, for (a) and (b)) are averaged. Now it really would help to know the future. There is only a small penalty for including irrelevant regressors, as their influence is swamped by the break. Except for the sample means, both feasible methods perform on a par with ARY. The infeasible device is best with loose selection, as was found theoretically.

9.2. Selection and Location of the Break

The design of the experiments allows for three locations of the break. Table 9 gives the mean square forecast errors for a break in mean and slope (b), listing three cases.

Table 9. Break in mean and slope (b). MSFE for different locations of the break.

			$T+1\|T$				$T+2\|T+1$				$T+3\|T+2$			
	ψ	**Where**	**INF**	**AVG**	**ARX**	**RWX**	**INF**	**AVG**	**ARX**	**RWX**	**INF**	**AVG**	**ARX**	**RWX**
$\alpha = 0$ ARY	$\psi(1)$	Relevant		54.42				50.41				6.75		
$\alpha = 0.1$	$\psi(1)$	Relevant	16.44	54.49	54.50	54.60	10.67	60.24	18.33	11.24	3.51	33.30	3.03	3.48
GUM	$\psi(1)$	Relevant	11.43	54.78	54.63	54.77	10.15	60.60	13.56	9.76	2.89	41.42	2.43	4.15
$\alpha = 0$ ARY	$\psi(1)$	All		54.42				50.41				6.75		
$\alpha = 0.1$	$\psi(1)$	All	18.32	54.49	54.50	54.60	11.19	61.32	18.71	11.70	3.32	33.28	2.89	3.61
GUM	$\psi(1)$	All	16.42	54.78	54.63	54.77	14.12	64.35	17.41	13.64	3.05	42.12	2.55	4.21
$\alpha = 0$ ARY	$\psi(1)$	Irrel.		1.15				1.19				1.18		
$\alpha = 0.1$	$\psi(1)$	Irrel.	3.19	1.36	1.26	1.31	2.86	2.59	2.55	2.80	1.82	2.04	1.80	1.89
GUM	$\psi(1)$	Irrel.	6.71	1.75	1.39	1.42	5.74	6.16	5.38	5.52	2.18	3.39	2.02	2.00
$\alpha = 0$ ARY	$\psi(2)$	Relevant		54.71				43.20				6.02		
$\alpha = 0.1$	$\psi(2)$	Relevant	7.90	54.86	54.82	54.98	4.84	61.25	12.40	5.43	2.64	37.40	2.51	3.87
GUM	$\psi(2)$	Relevant	7.60	55.05	54.94	55.17	6.73	58.20	10.67	6.58	2.68	40.60	2.47	4.31
$\alpha = 0$ ARY	$\psi(2)$	All		54.71				43.20				6.02		
$\alpha = 0.1$	$\psi(2)$	All	11.05	54.86	54.82	54.98	6.76	62.80	13.90	7.23	2.65	37.15	2.59	4.27
GUM	$\psi(2)$	All	16.45	55.05	54.94	55.17	13.99	64.88	17.65	13.72	3.02	42.09	2.71	4.41
$\alpha = 0$ ARY	$\psi(2)$	Irrel.		1.31				1.39				1.35		
$\alpha = 0.1$	$\psi(2)$	Irrel.	4.44	1.61	1.33	1.38	3.71	3.70	3.43	3.72	1.93	2.66	2.04	2.14
GUM	$\psi(2)$	Irrel.	10.46	1.97	1.48	1.53	8.90	9.47	8.59	8.78	2.36	4.11	2.28	2.26
$\alpha = 0$ ARY	$\psi(4)$	Relevant		54.98				39.74				5.66		
$\alpha = 0.1$	$\psi(4)$	Relevant	4.38	54.98	55.03	55.31	2.64	61.84	9.54	3.19	2.00	38.64	2.21	4.54
GUM	$\psi(4)$	Relevant	4.56	55.27	55.21	55.51	4.20	56.23	8.50	4.27	2.42	39.53	2.45	4.47
$\alpha = 0$ ARY	$\psi(4)$	All		54.98				39.74				5.66		
$\alpha = 0.1$	$\psi(4)$	All	8.45	54.98	55.03	55.31	5.27	63.59	11.82	5.71	2.31	38.55	2.55	5.00
GUM	$\psi(4)$	All	16.47	55.27	55.21	55.51	13.89	65.37	17.89	13.79	3.00	42.11	2.85	4.59
$\alpha = 0$ ARY	$\psi(4)$	Irrel.		1.43				1.52				1.49		
$\alpha = 0.1$	$\psi(4)$	Irrel.	5.20	1.82	1.39	1.45	4.09	4.37	3.92	4.23	1.92	3.09	2.16	2.25
GUM	$\psi(4)$	Irrel.	13.46	2.17	1.57	1.63	11.33	12.03	11.05	11.29	2.47	4.62	2.44	2.43

Break in relevant regressors ($\delta_R = -0.3, \lambda_R = 0.05, \delta_I = \delta, \lambda_I = \lambda$)

The break shows up in y through the relevant variables. Inclusion of irrelevant variables in the forecasting model is not costly relative to the impact of the break. Loose selection is preferred, because it includes more relevant variables. For $T+1|T$ selection has no impact because the break is not observed (except for known regressors). Including regressors in ARX and RWX gives a substantial improvement over ARY.

Break in irrelevant regressors ($\delta_I = -0.3, \lambda_I = 0.05, \delta_R = \delta, \lambda_R = \lambda$)

There is no break in y, so any inclusion of irrelevant variables is costly, as their break offsets the small estimated coefficients. The more irrelevant variables included, the stronger this effect. The autoregression in y is almost always preferred.

Break in all regressors ($\delta_R = \delta_I = -0.3, \lambda_R = \lambda_I = 0.05$)

The y variable is identical to that of a break in relevant variables only. Selection is now a trade-off between including variables that matter and help with forecasting, and irrelevant variables that make forecasts worse. Including regressors in ARX and RWX gives a substantial improvement over ARY.

9.3. Forecasting after the Break

We now dispense of INF for its infeasibility, and AVG because it has the highest MSFE in all experiments. Table 10 reports the ratio of the MSFE for all other devices to that of ARY. For the devices that forecast regressor values, results are reported after selection at 10%.

Table 10. Ratio of MSFE to that of MSFE_{ARY}. Selection at $\alpha = 0.1$ for ARX, RWX, RDX, and CAX.

	$T+2\|T+1$						$T+3\|T+2$						$T+4\|T+3$					
	ARX	**RWX**	**RDX**	**CAX**	**RWY**	**CAY**	**ARX**	**RWX**	**RDX**	**CAX**	**RWY**	**CAY**	**ARX**	**RWX**	**RDX**	**CAX**	**RWY**	**CAY**
	No break																	
$\psi(1)$	1.10	1.15	1.31	1.15	1.21	1.29	1.11	1.16	1.31	1.16	1.21	1.29	1.10	1.14	1.30	1.15	1.21	1.28
$\psi(2)$	1.00	1.04	1.24	1.05	1.14	1.19	1.02	1.07	1.29	1.08	1.15	1.22	1.01	1.06	1.25	1.06	1.15	1.21
$\psi(4)$	0.96	1.00	1.23	1.01	1.12	1.16	0.97	1.02	1.26	1.03	1.12	1.17	0.96	1.00	1.23	1.00	1.12	1.17
	Break in mean and slope (b) of irrelevant regressors																	
$\psi(1)$	2.14	2.35	3.30	2.33	1.21	1.29	1.53	1.61	1.75	1.64	1.21	1.29	1.20	1.25	1.35	1.25	1.21	1.28
$\psi(2)$	2.47	2.68	3.80	2.66	1.14	1.19	1.51	1.59	1.78	1.62	1.15	1.22	1.13	1.17	1.31	1.17	1.15	1.21
$\psi(4)$	2.58	2.79	3.90	2.76	1.12	1.16	1.45	1.51	1.73	1.53	1.12	1.17	1.07	1.11	1.29	1.11	1.12	1.17
	Break in mean of all regressors																	
$\psi(1)$	0.62	0.50	0.34	0.50	0.63	0.57	0.48	0.47	0.75	0.48	0.28	0.26	0.72	0.69	0.82	0.69	0.67	0.67
$\psi(2)$	0.57	0.42	0.25	0.42	0.69	0.62	0.50	0.58	1.19	0.60	0.37	0.34	0.79	0.84	0.92	0.85	0.85	0.85
$\psi(4)$	0.54	0.37	0.22	0.37	0.72	0.65	0.51	0.69	1.61	0.72	0.43	0.40	0.81	0.96	0.98	0.96	0.94	0.94
	Break in slope (a) of all regressors																	
$\psi(1)$	0.69	0.59	0.43	0.58	0.69	0.58	0.57	0.57	0.85	0.57	0.37	0.36	0.77	0.75	0.87	0.75	0.71	0.76
$\psi(2)$	0.64	0.51	0.34	0.50	0.73	0.63	0.58	0.66	1.29	0.68	0.48	0.46	0.82	0.87	0.98	0.86	0.87	0.92
$\psi(4)$	0.61	0.46	0.30	0.46	0.76	0.65	0.60	0.77	1.69	0.80	0.55	0.53	0.83	0.95	1.03	0.94	0.95	1.00
	Break in slope (b) of all regressors																	
$\psi(1)$	0.41	0.28	0.42	0.29	0.38	0.33	0.42	0.41	0.49	0.42	0.21	0.21	0.85	0.86	0.99	0.84	1.16	1.03
$\psi(2)$	0.36	0.21	0.59	0.22	0.44	0.38	0.43	0.54	0.70	0.57	0.28	0.28	0.87	1.03	1.04	0.98	1.29	1.17
$\psi(4)$	0.32	0.19	0.78	0.19	0.49	0.41	0.45	0.69	0.91	0.74	0.33	0.34	0.87	1.18	1.05	1.11	1.35	1.24
	Break in mean and slope (a) of all regressors																	
$\psi(1)$	0.83	0.78	0.75	0.78	0.86	0.87	0.88	0.91	1.11	0.92	0.79	0.82	0.99	1.01	1.14	1.01	1.00	1.05
$\psi(2)$	0.76	0.69	0.67	0.69	0.86	0.87	0.87	0.94	1.31	0.95	0.86	0.88	0.97	1.01	1.17	1.01	1.06	1.10
$\psi(4)$	0.73	0.65	0.63	0.64	0.87	0.87	0.85	0.95	1.42	0.96	0.88	0.91	0.93	1.00	1.18	1.00	1.08	1.10
	Break in mean and slope (b) of all regressors																	
$\psi(1)$	0.37	0.23	0.39	0.25	0.35	0.32	0.43	0.53	0.67	0.60	0.21	0.22	0.86	1.09	1.07	1.03	1.64	1.44
$\psi(2)$	0.32	0.17	0.55	0.18	0.42	0.37	0.43	0.71	0.94	0.82	0.26	0.27	0.83	1.25	1.04	1.17	1.66	1.51
$\psi(4)$	0.30	0.14	0.71	0.16	0.46	0.40	0.45	0.88	1.19	1.04	0.30	0.31	0.82	1.41	1.02	1.30	1.67	1.55
	Average over all breaks in all regressors																	
$\psi(1)$	0.58	0.48	0.47	0.48	0.58	0.54	0.56	0.58	0.77	0.60	0.37	0.37	0.84	0.88	0.98	0.87	1.04	0.99
$\psi(2)$	0.53	0.40	0.48	0.40	0.63	0.58	0.56	0.69	1.08	0.72	0.45	0.45	0.86	1.00	1.03	0.97	1.15	1.11
$\psi(4)$	0.50	0.36	0.53	0.36	0.66	0.60	0.57	0.80	1.37	0.85	0.50	0.50	0.85	1.10	1.05	1.06	1.20	1.17

When there is no break, only ARX is able to gain on ARY, and then only for the design with significant regressors (but stricter selection would help; see Table 8). Otherwise, and always for the break in irrelevant variables only, the AR(1) in y has the smallest mean square forecast error. This matches an oft-found outcome. This model is misspecified, ignoring all information from the exogenous regressors, but misspecification need not entail forecast failure. Indeed, the costs of forecasting the exogenous regressors can outweigh their inclusion. However, the DGP design is also an AR(1) in y so this forecasting device has the advantage of correctly specifying the dynamics. It may not perform so well if the DGP contains more complex dynamics.

The AR(1) in y performs poorly when relevant regressors break. Now we see substantial gains in Table 10 from modeling the regressors, even shortly after the break has finished (the break is active for $T+1$ and $T+2$).

Device RDX improves on RWX when the process shifts towards a unit root, but not otherwise. Cardt behaves quite similar to the random walk forecasts in this DGP: CAX is close to RWX in most cases. Cardt on y is usually a small improvement on RWY in the cases with a break.

The AR(1) for x always improves on ARY in the cases with break. In the first period with an observed break, $T+2$, it is the worst of the methods that forecast regressors, while in subsequent periods it is the best of these. But note that at $T+3$ the naive random walk forecast of y and Cardt are better still.

9.4. Is Selection Costly When Forecasting?

Comparing selection to using the GUM to forecast regressors, we find that selection is always advantageous. Table 11 gives the average MSFE ratio relative to ARY, where the average is taken over the three noncentrality settings, and different break cases. The top panel of the table combines cases where there is no change in y, either because nothing breaks, or for the break in mean and slope for irrelevant variables only. In that case ARY tends to dominate, so tight selection is advantageous. The exception is highly significant regressors in a stationary setting.

Table 11. Ratio of MSFE to that of MSFE_{ARY}. Average over noncentralities.

	$T+2\|T+1$					$T+3\|T+2$					$T+4\|T+3$				
	INF	**ARX**	**RWX**	**RDX**	**CAX**	**INF**	**ARX**	**RWX**	**RDX**	**CAX**	**INF**	**ARX**	**RWX**	**RDX**	**CAX**
	No break in y: no break and break in irrelevant variables														
$\alpha = 0.01$	1.13	1.15	1.22	1.55	1.22	0.99	1.07	1.13	1.26	1.13	0.93	1.01	1.04	1.16	1.04
$\alpha = 0.05$	1.52	1.47	1.59	2.14	1.57	1.14	1.21	1.27	1.45	1.28	0.99	1.05	1.10	1.25	1.10
$\alpha = 0.1$	1.78	1.71	1.84	2.46	1.81	1.21	1.26	1.33	1.52	1.33	1.03	1.08	1.12	1.29	1.12
GUM	3.69	3.55	3.64	4.17	3.61	1.46	1.41	1.42	1.60	1.41	1.21	1.17	1.19	1.41	1.19
DGP	0.82	0.92	0.96	1.12	0.96	0.83	0.93	0.97	1.14	0.97	0.82	0.93	0.96	1.13	0.96
	Break in y: break in all variables														
$\alpha = 0.01$	0.40	0.63	0.52	0.52	0.52	0.59	0.62	0.68	0.94	0.70	0.82	0.87	0.99	1.00	0.97
$\alpha = 0.05$	0.31	0.56	0.43	0.48	0.44	0.54	0.56	0.67	1.02	0.70	0.82	0.85	0.99	1.01	0.96
$\alpha = 0.1$	0.30	0.54	0.41	0.49	0.42	0.55	0.56	0.69	1.07	0.72	0.83	0.85	0.99	1.02	0.97
GUM	0.38	0.54	0.44	0.64	0.43	0.67	0.65	0.79	1.26	0.83	0.96	0.91	1.00	1.09	0.98
DGP	0.17	0.44	0.29	0.41	0.30	0.36	0.40	0.61	1.11	0.66	0.64	0.72	1.02	0.86	0.97

The bottom panel of Table 11 averages over the five cases where all variables break. There we often see a U-shaped effect of selection, with a loose selection best. This is particularly so at $T+2|T+1$, as was found in the theoretical results.

The bottom row in each panel of Table 11 gives the result when the specification of the DGP is known but its parameters need estimated. The entries under INF have the most information: the DGP as well as the future values of the regressors. Moving to the other columns shows the cost of not knowing the latter.

9.5. Forecast Combinations

Many investigations of forecasting have shown that combined forecasts can outperform the individual forecasts. The main candidates here are ARX in combination with a random walk style forecast of y. Although there are many other possibilities, we restrict ourselves to:

APOOL (ARX + RWY)/2;

CPOOL (ARX + CAY)/2.

In both cases ARX is used in the model that is selected from the GUM at 10%.

To summarize the results, we consider again the MSFE relative to ARY, with a three-way average across noncentralities, break types and horizons $T+2, T+3, T+4$. Table 12 illustrates that in this setting pooling can be advantageous as well. It is even competitive with the infeasible device.

Table 12. Ratio of MSFE to that of MSFE_{ARY}. Selection at $\alpha = 0.1$. Average over noncentralities and horizons $T+2, \ldots, T+4$. Lowest two in bold (excluding INF).

	INF	AVG	ARX	RWX	RDX	CAX	RWY	CAY	APOOL	CPOOL	ARY
No break	0.99	1.22	1.03	1.07	1.27	1.07	1.16	1.22	**0.96**	**1.00**	**1.00**
Break irrelevant	1.69	1.99	1.68	1.78	2.25	1.77	1.16	1.22	**1.13**	1.20	**1.00**
All breaks	0.56	2.93	**0.65**	0.70	0.86	0.70	0.73	0.70	0.73	**0.58**	1.00
Sum	3.24	6.14	3.36	3.55	4.38	3.54	3.05	3.14	**2.82**	**2.78**	3.00

9.6. Summary of the Simulation Results

We can infer some general results from the experiments. First, using the in-sample mean to forecast the exogenous regressors is always dominated by other approaches.

Next, when the break occurs out of sample, so forecasts are computed for $T+1$, all methods struggle, and incorporating regressors is worse than simply using the AR(1) for y. Moving to the case when the break occurs in sample, so the forecasts are computed for $T+2$ when the break occurs at $T+1$, the random walk forecasts of the regressors is preferred when the break occurs in the relevant or all regressors. Looser significance levels tend to do well here. If the breaks occur in the irrelevant regressors, including even one can already be poisonous, and the AR(1) in y performs best.

There are substantial differences in the forecast performance of the two robust devices RWX and RDX. The former is the random walk for the regressor, and works best, except if the break drives the process towards a unit root. In that case, the differenced AR(1) for x gives a higher weight to the previous value. However, when the type of break is unknown, represented by the average performance here, the simple random walk dominates.

Table 12, rather arbitrarily, averages over all experiments and horizons. It shows that pooling provides some protection against different states of nature, just inching ahead of the autoregression in y. After that come the methods that ignore regressors, followed by using an AR(1), random walk, or Cardt, to forecast the regressors. However, if we know that a break has happened in the regressors, we should switch to modeling them, at least until the break is out of the system again.

The variation in MSFEs across α is very small for intermediate values of α relative to the variation in MSFEs across break types and DGP designs. For moderate α the selection significance level does not have a large impact on forecast performance. This is an encouraging finding showing that forecast performance is relatively unaffected by the precise choice of significance level for selection when using *Autometrics*, despite a range of noncentralities and numbers of relevant and irrelevant exogenous variables.

10. Conclusions

This paper investigates the choice of significance level and its associated critical value when selecting forecasting models, both analytically in a static bivariate setting where

there are location shifts at the forecast origin, and in more general simulation experiments. The theory suggests that variables should be retained if their noncentralities exceed 1, which translates to $c_\alpha^2 = 2$ at the boundary. This result holds regardless of whether location shifts affect the variable about which a retention decision is made. Undertaking selection at such loose significance levels implies that fewer relevant variables will be excluded when they contribute to forecast accuracy, but that more variables will be retained by chance because they happen to be in a draw that results in statistical significance at the proposed critical value. Although retaining irrelevant variables that are subject to location shifts usually worsens forecast performance, their coefficient estimates will be driven towards zero when updating estimates as the horizon moves forward.

Although the static design is simple, it produces several generic analytical results. Those results hold regardless of whether the regressors are contemporaneous or lagged, although the timing of location shifts is fundamental. Dynamics will slow adjustment to new equilibria, but this would not change the essence of the results. The inflation forecasts illustrated the analytic results, with a loose selection significance level of 16% being preferred for both the known regressors and the random walk forecasts for unknown regressors case.

The simulation evidence examines a wide range of experimental designs and despite the disparate outcomes, they provide some guidance for forecasting. The ideal scenario is obviously to have complete knowledge of the DGP, such that the empirical modeller knows the number and magnitude of both relevant and irrelevant regressors, and their future values, and hence whether and where breaks are likely to occur. In practice, no-one has the benefit of omniscience, and once the future values of regressors need to be forecast, selecting from a GUM that nests the DGP may cost little, relative to knowing the precise specification of the DGP.

The simulation results suggest that if the model is being used primarily for one-step-ahead forecasting with the aim of minimizing MSFE, selection at looser than standard selection significance levels may well help, and doing so will rarely hinder forecast performance. The results provide some support for selecting models at around 10% when there are approximately 15 regressors, many of which are irrelevant. This is close to the 16% derived theoretically in this paper when the number of irrelevant regressors is small. The simulation results also highlight the degree of complexity in pinning down the optimal selection rule for forecasting, with results depending on all aspects of the experimental design. A take-away for the forecaster is that pooling works well across many settings, suggesting a combination of a robust device which minimizes systematic bias and model-based forecast based on univariate methods as a good insurance policy. Moreover, methods that did not nest the DGP, such as the direct AR(1) forecast of the dependent variable and Cardt, also performed well, both matching commonly found empirical outcomes. However, if we know that a break has happened, one-step forecasts are improved by incorporating forecasts of the regressors.

Author Contributions: Conceptualization, J.L.C., J.A.D. and D.F.H.; Methodology, J.L.C., J.A.D. and D.F.H.; Software, J.A.D.; Formal Analysis, J.L.C., J.A.D. and D.F.H.; Writing and Original Draft Preparation, J.L.C., J.A.D. and D.F.H.; Writing Review and Editing, J.L.C., J.A.D. and D.F.H. All authors have read and agreed to the published version of the manuscript.

Funding: Financial support from the Robertson Foundation (award 9907422), the Institute for New Economic Thinking (grant 20029822), and the ERC (grant 694262, DisCont) is gratefully acknowledged.

Data Availability Statement: Data available from stated sources.

Acknowledgments: We thank participants at the 2018 International Symposium of Forecasting, the 7th Rhenish Multivariate Time Series Econometrics Meeting in Koblenz, the 20th OxMetrics Users Conference, and the 2nd Forecasting at Central Banks Conference at the Bank of England for helpful comments, as well as members of the Economics Department Econometrics Lunch group at Oxford University, Michael P. Clements, Andrew B. Martinez, Felix Pretis, and Sophocles Mavroeidis. We thank Michael McCracken for suggesting comparisons with bagging which we will investigate in

future research. We are especially grateful to Neil Ericsson and two anonymous referees for their careful reading and many helpful comments.

Conflicts of Interest: Doornik and Hendry have developed Autometrics, which is included in the OxMetrics software package, and have a share in the returns.

Appendix A. Analytic Calculations

Appendix A.1

Derivations for the equations reported in Section 3.

The DGP given in (1)–(3) results in

$$\sqrt{T}\begin{pmatrix} \widehat{\beta}_1 - \beta_1 \\ \widehat{\beta}_2 - \beta_2 \end{pmatrix} \sim \mathrm{N}_2\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \frac{\sigma_\epsilon^2}{\sigma_{11}^2\sigma_{22}^2(1-\rho^2)}\begin{pmatrix} \sigma_{22}^2 & -\rho\sigma_{11}\sigma_{22} \\ -\rho\sigma_{11}\sigma_{22} & \sigma_{11}^2 \end{pmatrix}\right],$$

with:

$$\sqrt{T}(\mu_y - \widehat{\mu}_y) \sim \mathrm{N}\left[0, \sigma_\epsilon^2\right],$$

where we subsequently set $\sigma_{11} = \sigma_{22} = 1$ without loss of generality.

M_2 in (6) partials out $x_{2,t}$. From (2) we can write in deviations from means for $t = 1, \ldots, T$:

$$x_{2,t} - \mu_2 = \rho(x_{1,t} - \mu_1) + e_t,$$

such that $e_t = \eta_{2,t} - \rho\eta_{1,t}$, so $\gamma_1 = (\beta_1 + \beta_2\rho)$ and $\phi_0 = \mu_y - \gamma_1\mu_1$. Hence M_2 is:

$$\begin{aligned} y_t &= \mu_y + (\beta_1 + \beta_2\rho)(x_{1,t} - \mu_1) + \beta_2 e_t + \epsilon_t \\ &= \gamma_0 + \gamma_1(x_{1,t} - \mu_1) + \nu_t, \end{aligned}$$

with $\gamma_0 = \mu_y$. The error for M_2 is given by:

$$\nu_t = \beta_2(\eta_{2,t} - \rho\eta_{1,t}) + \epsilon_t,$$

where

$$\sigma_\nu^2 = \sigma_\epsilon^2 + \beta_2^2\left(1 - \rho^2\right) = \sigma_\epsilon^2\left(1 + T^{-1}\psi_\beta^2\right) \geqq \sigma_\epsilon^2. \tag{A1}$$

Also

$$\sqrt{T}\begin{pmatrix} \widetilde{\gamma}_0 - \gamma_0 \\ \widetilde{\gamma}_1 - \gamma_1 \end{pmatrix} \sim \mathrm{N}_2\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \sigma_\nu^2\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right].$$

Appendix A.2

Derivations for the equations reported in Section 4.

The one-step-ahead forecast error from M_1 is:

$$\begin{aligned} \widehat{\epsilon}_{T+1|T} &= y_{T+1} - \widehat{y}_{T+1|T} \\ &= (\mu_y - \widehat{\mu}_y) + \left(\beta_1 - \widehat{\beta}_1\right)(x_{1,T+1} - \mu_1) + \left(\beta_2 - \widehat{\beta}_2\right)(x_{2,T+1} - \mu_2) + \epsilon_{T+1}. \end{aligned}$$

When there are no breaks, the parameter estimates are unbiased, $\mathrm{E}\left[\widehat{\epsilon}_{T+1|T}\right] = 0$ so the MSFE of M_1 is:

$$\mathrm{E}\left[\widehat{\epsilon}_{T+1|T}^2\right] = \sigma_\epsilon^2\left(1 + \frac{1}{T} + \frac{2}{T(1-\rho^2)} - \frac{2\rho^2}{T(1-\rho^2)}\right) = \sigma_\epsilon^2\left(1 + \frac{3}{T}\right).$$

The one-step-ahead forecast error from M_2 in which $x_{2,t}$ is omitted is:

$$\begin{aligned} \widetilde{\epsilon}_{T+1|T} &= y_{T+1} - \widetilde{y}_{T+1|T} \\ &= \beta_2\eta_{2,T+1} + \epsilon_{T+1} + (\gamma_0 - \widetilde{\gamma}_0) + (\beta_1 - \widetilde{\gamma}_1)\eta_{1,T+1}. \end{aligned}$$

Therefore, despite the misspecification, $\mathrm{E}\left[\widetilde{\epsilon}_{T+1|T}\right] = 0$ and the MSFE is:

$$\mathrm{E}\left[\widetilde{\epsilon}^2_{T+1|T}\right] = \mathrm{E}\left[(\beta_2\eta_{2,T+1} + \epsilon_{T+1} + (\gamma_0 - \widetilde{\gamma}_0) + (\beta_1 - \widetilde{\gamma}_1)\eta_{1,T+1})^2\right] = \sigma_\nu^2\left(1 + \frac{2}{T}\right).$$

Appendix A.3

Derivations for the equations reported in Section 5.2.

The regression equation itself stays constant so:

$$y_{T+1} = (\mu_y + \beta_2\delta) + \beta_1(x_{1,T+1} - \mu_1) + \beta_2(x_{2,T+1} - \mu_2 - \delta) + \epsilon_{T+1}. \tag{A2}$$

Consequently, using $\widehat{\beta}_0 = \mu_y - \widehat{\beta}_1\mu_1 - \widehat{\beta}_2\mu_2$ to match the formulation of M_2, the forecast for M_1 is:

$$\overline{\widehat{y}}_{T+1|T+1} = \mu_y + \widehat{\beta}_2\delta + \widehat{\beta}_1(x_{1,T+1} - \mu_1) + \widehat{\beta}_2(x_{2,T+1} - \mu_2 - \delta),$$

and the one-step-ahead forecast error for M_1 is:

$$\begin{aligned}\overline{\widehat{\epsilon}}_{T+1|T+1} &= y_{T+1} - \overline{\widehat{y}}_{T+1|T+1} \\ &= \left(\beta_2 - \widehat{\beta}_2\right)\delta + \left(\beta_1 - \widehat{\beta}_1\right)\eta_{1,T+1} + \left(\beta_2 - \widehat{\beta}_2\right)\eta_{2,T+1} + \epsilon_{T+1},\end{aligned}$$

and a one-step-ahead MSFE of:

$$\mathrm{E}\left[\overline{\widehat{\epsilon}}^2_{T+1|T+1}\right] = \sigma_\epsilon^2\left(1 + \frac{\delta^2 + 2 - \rho}{T(1-\rho^2)}\right).$$

Next consider the one-step-ahead forecast for M_2, given $\gamma_0 = \mu_y$ and $\gamma_1 = (\beta_1 + \beta_2\rho)$:

$$\overline{\widetilde{y}}_{T+1|T+1} = \widetilde{\gamma}_0 + \widetilde{\gamma}_1(x_{1,T+1} - \mu_1).$$

The one-step-ahead forecast error is given by:

$$\begin{aligned}\overline{\widetilde{\epsilon}}_{T+1|T+1} &= y_{T+1} - \overline{\widetilde{y}}_{T+1|T+1} \\ &= \beta_2\delta + (\gamma_0 - \widetilde{\gamma}_0) + (\gamma_1 - \widetilde{\gamma}_1)\eta_{1,T+1} - \beta_2\rho\eta_{1,T+1} + \beta_2\eta_{2,T+1} + \epsilon_{T+1},\end{aligned}$$

and the one-step-ahead MSFE for M_2 is:

$$\mathrm{E}\left[\overline{\widetilde{\epsilon}}^2_{T+1|T+1}\right] = \sigma_\epsilon^2 + \beta_2^2\left(1 - \rho^2 + \delta^2\right) + 2T^{-1}\sigma_\nu^2.$$

Appendix A.4

Derivations for the equations reported in Section 5.4.

For $\widehat{\beta}_0 = \mu_y - \widehat{\beta}_1\mu_1 - \widehat{\beta}_2\mu_2$, replacing the unknown $x_{i,T+1}$ by μ_i leads to forecasting y_{T+1} by the in-sample mean:

$$\widehat{\widehat{y}}_{T+1|T} = \mu_y,$$

so the forecast error for M_1 is:

$$\begin{aligned}\widehat{\widehat{\epsilon}}_{T+1|T} &= y_{T+1} - \widehat{\widehat{y}}_{T+1|T} \\ &= \beta_2\delta + \beta_1\eta_{1,T+1} + \beta_2\eta_{2,T+1} + \epsilon_{T+1},\end{aligned}$$

and the forecast error bias is:

$$\mathrm{E}\left[\widehat{\widehat{\epsilon}}_{T+1|T}\right] = \beta_2\delta.$$

The MSFE_1 is:

$$\mathrm{E}\left[\widehat{\widetilde{\epsilon}}^2_{T+1|T}\right] = \beta_1^2 + \beta_2^2\left(1+\delta^2\right) + 2\rho\beta_1\beta_2 + \sigma_\epsilon^2.$$

Parameter estimation adds terms of $\mathrm{O}_p\left(T^{-1}\right)$.

Similarly, for M_2, from (6) forecasting $x_{1,T+1}$ by μ_1 leads to:

$$\widetilde{\widetilde{y}}_{T+1|T} = \mu_y,$$

and hence for 'known' μ_y the forecast error is:

$$\widetilde{\widetilde{\epsilon}}_{T+1|T} = \beta_2\delta + \beta_1\eta_{1,T+1} + \beta_2\eta_{2,T+1} + \epsilon_{T+1} = \widehat{\widetilde{\epsilon}}_{T+1|T},$$

with

$$\mathrm{E}\left[\widetilde{\widetilde{\epsilon}}_{T+1|T}\right] = \beta_2\delta,$$

and MSFE_2 is given by (23). Hence, ignoring $\mathrm{O}_p\left(T^{-1}\right)$ terms, $\text{MSFE}_2 = \text{MSFE}_1$.

Appendix A.5

Derivations for the equations reported in Section 5.5.

From (A2) the regression equation for y_{T+1} can also be written as:

$$y_{T+1} = \left(\mu_y + \beta_2\delta\right) + \beta_1\Delta x_{1,T+1} + \beta_2(\Delta x_{2,T+1} - \delta) + \epsilon_{T+1} + \beta_1\eta_{1,T} + \beta_2\eta_{2,T}.$$

Furthermore, the forecast for M_1 using (24) and (25) is:

$$\overline{\widetilde{y}}_{T+1|T} = \mu_y + \widehat{\beta}_1(x_{1,T} - \mu_1) + \widehat{\beta}_2(x_{2,T} - \mu_2),$$

so the forecast error for M_1 is:

$$\begin{aligned}\overline{\widetilde{\epsilon}}_{T+1|T} &= y_{T+1} - \overline{\widetilde{y}}_{T+1|T} \\ &= \beta_2\delta + \beta_1\Delta x_{1,T+1} + \beta_2(\Delta x_{2,T+1} - \delta) + \left(\beta_1 - \widehat{\beta}_1\right)\eta_{1,T} + \left(\beta_2 - \widehat{\beta}_2\right)\eta_{2,T} + \epsilon_{T+1}.\end{aligned}$$

Consequently, neglecting the small impact of $\eta_{i,T}$ on $\beta_i - \widehat{\beta}_i$:

$$\mathrm{E}\left[\overline{\widetilde{\epsilon}}_{T+1|T}\right] = \beta_2\delta,$$

and hence MSFE_1 is:

$$\mathrm{E}\left[\overline{\widetilde{\epsilon}}^2_{T+1|T}\right] = 2\beta_1^2 + \beta_2^2\left(2+\delta^2\right) + 4\rho\beta_1\beta_2 + \sigma_\epsilon^2\left(1+2T^{-1}\right).$$

Next, we compute the equivalent bias and MSFE for M_2, noting $\gamma_1 = \beta_1 + \beta_2\rho$, so that the forecast is given by:

$$\widetilde{\widetilde{y}}_{T+1|T} = \widetilde{\gamma}_0 + \widetilde{\gamma}_1(x_{1,T} - \mu_1).$$

As $\widetilde{\gamma}_0 = \gamma_0 = \mu_y$, the forecast error for M_2 using the random walk is:

$$\begin{aligned}\widetilde{\widetilde{\epsilon}}_{T+1|T} &= y_{T+1} - \widetilde{\widetilde{y}}_{T+1|T} \\ &= \beta_2\delta + \beta_1\Delta\eta_{1,T+1} + \beta_2\Delta\eta_{2,T+1} + \epsilon_{T+1} + (\beta_1 - \widetilde{\gamma}_1)\eta_{1,T} + \beta_2\eta_{2,T},\end{aligned}$$

where, as before:

$$\mathrm{E}\left[\widetilde{\widetilde{\epsilon}}_{T+1|T}\right] = \beta_2\delta.$$

Neglecting the small impact of $\eta_{1,T}$ on $\widetilde{\gamma}_1$ the MSFE for M_2 is:

$$\mathrm{E}\left[\widetilde{\widetilde{\epsilon}}^2_{T+1|T}\right] = 2\beta_1^2 + \beta_2^2\left(3+\rho^2+\delta^2\right) + 4\rho\beta_1\beta_2 + \sigma_\epsilon^2\left(1+T^{-1}+T^{-2}\psi_\beta^2\right).$$

Appendix A.6

Derivations for the equations reported in Section 6.2.

The conditional DGP for the forecast observation is:

$$\begin{aligned} y_{T+1} &= \beta_0 + \beta_1 x_{1,T+1} + \beta_2 x_{2,T+1} + \epsilon_{T+1} \\ &= (\mu_y + \beta_2\delta) + \beta_1(x_{1,T+1} - \mu_1) + \beta_2(x_{2,T+1} - \mu_2 - \delta) + \epsilon_{T+1}, \end{aligned} \tag{A3}$$

where the in-sample mean μ_y is shifted to $(\mu_y + \beta_2\delta)$ at T. Sample calculations will be altered as now $\mathrm{E}[\overline{x}_2] = \mu_2 + T^{-1}\delta$ from:

$$\overline{x}_2 = \frac{1}{T}\sum_{t=1}^{T} x_{2,t} = \mu_2 + T^{-1}\delta + \overline{\eta}_2,$$

and neglecting terms of T^{-2} or smaller:

$$(\sigma^*_{22})^2 \approx \sigma^2_{22} + T^{-1}\delta^2,$$

with $\sigma^*_{12} = \sigma_{12}$ implying that:

$$\rho^* = \frac{\sigma_{12}}{\sigma_{11}\sigma^*_{22}}.$$

The intercept is again included with $\widehat{\beta}_0 = \mu_y - \widehat{\beta}_1\mu_1 - \widehat{\beta}_2\mu_2$ to match the formulation of M_2.

$$\widehat{\widehat{y}}_{T+1|T+1} \approx \widehat{\beta}_0 + \widehat{\beta}_1\mu_1 + \widehat{\beta}_2\Big(\mu_2 + T^{-1}\delta\Big) = \mu_y + \widehat{\beta}_2 T^{-1}\delta,$$

and hence neglecting terms of T^{-2} or smaller, the forecast error for M_1 is:

$$\begin{aligned} \widehat{\widehat{\epsilon}}_{T+1|T+1} &= y_{T+1} - \widehat{\widehat{y}}_{T+1|T+1} \\ &\approx \beta_2\delta\Big(1 - T^{-1}\Big) + \beta_1\eta_{1,T+1} + \beta_2\eta_{2,T+1} + \epsilon_{T+1}, \end{aligned}$$

so the forecast error bias is given by:

$$\mathrm{E}\Big[\widehat{\widehat{\epsilon}}_{T+1|T+1}\Big] \approx \beta_2\delta\Big(1 - T^{-1}\Big).$$

The MSFE for M_1 is:

$$\mathrm{E}\Big[\widehat{\widehat{\epsilon}}^2_{T+1|T+1}\Big] \approx \beta_2^2\delta^2\Big(1 - T^{-1}\Big)^2 + \beta_1^2 + \beta_2^2 + \sigma^2_\epsilon.$$

Omitting x_2 from the forecasting equation leads to a forecast error of:

$$\begin{aligned} \widehat{\widetilde{\epsilon}}_{T+1|T+1} &= y_{T+1} - \widehat{\widetilde{y}}_{T+1|T+1} \\ &\approx \beta_2\delta + (\gamma_0 - \widetilde{\gamma}_0) + (\gamma_1 - \widetilde{\gamma}_1)\eta_{1,T+1} + v_{T+1}, \end{aligned}$$

with an MSFE for M_2 given by:

$$\mathrm{E}\Big[\widehat{\widetilde{\epsilon}}^2_{T+1|T+1}\Big] \approx \beta_2^2\delta^2 + \sigma^2_\epsilon + \sigma^2_v\Big(1 + \frac{2}{T}\Big),$$

where σ^2_v is given in (A1).

Appendix A.7

Derivations for the equations reported in Sections 6.4 and 6.5.

Following a similar strategy as the previous analysis, including the intercept for comparability where $\widehat{\beta}_0 = \mu_y - \widehat{\beta}_1\mu_1 - \widehat{\beta}_2\mu_2$, then the forecast for M_1 is:

$$\overline{\widehat{y}}_{T+1|T+1} = \widehat{\beta}_0 + \widehat{\beta}_1\overline{\widetilde{x}}_{1,T+1|T} + \widehat{\beta}_2\overline{\widetilde{x}}_{2,T+1|T} = \mu_y + \widehat{\beta}_2\delta + \widehat{\beta}_1\eta_{1,T+1} + \widehat{\beta}_2\eta_{2,T+1},$$

so that the forecast error for M_1 is:

$$\begin{aligned}\overline{\widehat{\epsilon}}_{T+1|T} &= y_{T+1} - \overline{\widehat{y}}_{T+1|T} \\ &= \left(\beta_2 - \widehat{\beta}_2\right)\delta + \beta_1\Delta\eta_{1,T+1} + \beta_2\Delta\eta_{2,T+1} + \epsilon_{T+1} + \left(\beta_1 - \widehat{\beta}_1\right)\eta_{1,T} + \left(\beta_2 - \widehat{\beta}_2\right)\eta_{2,T},\end{aligned}$$

with $\mathrm{E}\left[\overline{\widehat{\epsilon}}_{T+1|T}\right] = 0$ when the parameter estimates are unbiased. The MSFE for M_1 is:

$$\mathrm{E}\left[\overline{\widehat{\epsilon}}^2_{T+1|T}\right] = 2\left(\beta_1^2 + \beta_2^2 + 2\rho\beta_1\beta_2\right) + \sigma_\epsilon^2\left(1 + T^{-1}\left(2 + \frac{\delta^2}{(1-\rho^2)}\right)\right).$$

Next we compute the random walk forecast for M_2 so $\gamma_1 = \beta_1 + \beta_2\rho$ and $\gamma_0 = \mu_y$, leading to the forecast given by:

$$\overline{\widetilde{y}}_{T+1|T} = \widetilde{\gamma}_0 + \widetilde{\gamma}_1(x_{1,T} - \mu_1),$$

and the forecast error for M_2 is:

$$\begin{aligned}\overline{\widetilde{\epsilon}}_{T+1|T} &= y_{T+1} - \overline{\widetilde{y}}_{T+1|T} \\ &= \beta_2\delta + \beta_1\Delta\eta_{1,T+1} + \beta_2\Delta\eta_{2,T+1} + \epsilon_{T+1} + (\beta_1 - \widetilde{\gamma}_1)\eta_{1,T} + \beta_2\eta_{2,T},\end{aligned}$$

which is now biased for $\beta_2\delta \neq 0$. The MSFE for M_2 is:

$$\mathrm{E}\left[\overline{\widetilde{\epsilon}}^2_{T+1|T}\right] = 2\beta_1^2 + \beta_2^2\left(\delta^2 + 1 + \rho^2\right) + 4\rho\beta_1\beta_2 + \sigma_\epsilon^2\left(1 + T^{-1} + T^{-2}\psi_\beta^2\right).$$

From (12):

$$\mathrm{MSFE}_3 = \mathrm{MSFE}_1 + \left(1 - \mathrm{p}_\alpha(\psi_\beta)\right)\left[\beta_2^2(\delta^2 + \rho^2 - 1) + \sigma_\epsilon^2\left(\tfrac{-\delta^2}{T(1-\rho^2)} - T^{-1} + T^{-2}\psi_\beta^2\right)\right]. \quad \text{(A4)}$$

Appendix B

Table A1. Ratio of MSFE to that of MSFE_1. $T = 100$, otherwise as Table 3.

Model	MSFE Relative to MSFE_1				
	$\psi_\beta^2 = 0$	$\psi_\beta^2 = 1$	$\psi_\beta^2 = 4$	$\psi_\beta^2 = 9$	$\psi_\beta^2 = 16$
Sections 4.1 and 4.2 No shift with known future regressors					
$\alpha = 0$ (M_2)	0.990	1.000	1.030	1.079	1.149
$\alpha = 0.001$	0.990	1.000	1.026	1.048	1.035
$\alpha = 0.05$	0.991	1.000	1.014	1.012	1.003
$\alpha = 0.16$	0.992	1.000	1.008	1.004	1.001
Sections 5.2 and 5.3 Out-of-sample shift with known future regressors					
$\alpha = 0$ (M_2)	0.827	1.008	1.551	2.457	3.724
$\alpha = 0.001$	0.827	1.008	1.497	1.895	1.651
$\alpha = 0.05$	0.836	1.007	1.267	1.217	1.056
$\alpha = 0.16$	0.855	1.005	1.152	1.081	1.013
Section 5.4 Out-of-sample shift with mean forecast of future regressors					
$\alpha = 0$ (M_2)	1.000	1.000	1.000	1.000	1.000
$\alpha = 0.001$	1.000	1.000	1.000	1.000	1.000
$\alpha = 0.05$	1.000	1.000	1.000	1.000	1.000
$\alpha = 0.16$	1.000	1.000	1.000	1.000	1.000

Table A1. *Cont.*

Model	MSFE Relative to $MSFE_1$				
	$\psi_\beta^2 = 0$	$\psi_\beta^2 = 1$	$\psi_\beta^2 = 4$	$\psi_\beta^2 = 9$	$\psi_\beta^2 = 16$
Section 5.5 Out-of-sample shift with random walk forecast of future regressors					
$\alpha = 0$ (M_2)	0.997	1.002	1.013	1.024	1.033
$\alpha = 0.001$	0.997	1.002	1.012	1.015	1.008
$\alpha = 0.05$	0.997	1.002	1.006	1.004	1.001
$\alpha = 0.16$	0.997	1.001	1.004	1.001	1.000
Sections 6.2 and 6.3 In-sample shift with mean forecast of future regressors					
$\alpha = 0$ (M_2)	1.010	1.009	1.008	1.007	1.007
$\alpha = 0.001$	1.010	1.009	1.008	1.005	1.002
$\alpha = 0.05$	1.010	1.008	1.004	1.001	1.000
$\alpha = 0.16$	1.008	1.006	1.002	1.000	1.000
Sections 6.4 and 6.5 In-sample shift with random walk forecast of future regressors					
$\alpha = 0$ (M_2)	0.931	0.994	1.155	1.386	1.661
$\alpha = 0.001$	0.931	0.994	1.140	1.237	1.158
$\alpha = 0.05$	0.934	0.995	1.075	1.058	1.014
$\alpha = 0.16$	0.942	0.996	1.043	1.021	1.003

Notes

[1] Clements and Hendry (1993) argue that the generalized forecast error second moment should be used to evaluate forecast performance instead of MSFE. In this case the results would be equivalent, because we focus on one-step-ahead forecasts.

[2] UK quarterly consumer price index (CPI) is given by ONS series D7BT, which is the quarterly average of the monthly index. Annual inflation percentage is defined as $\pi_t = 100\Delta_4 \log \mathrm{D7BT}_t$. UK Unemployment is the quarterly average of ONS series MGUK, LFS ILO unemployment rate (UK, All, Aged 16 and over, %, NSA).

[3] Intermediate alternatives such as sub-sample estimation, recursive or rolling estimation could also be used.

[4] Castle et al. (2012) demonstrate the ability of IIS to detect breaks in the form of location shifts at any point in the sample.

References

Akaike, Hirotogu. 1973. Information theory and an extension of the maximum likelihood principle. In *Second International Symposium of Information Theory*. Edited by Boris N. Petrov and Frigyes Csaki. Budapest: Akademiai Kiado, pp. 267–81.

Bontemps, Christophe, and Grayham E. Mizon. 2003. Congruence and encompassing. In *Econometrics and the Philosophy of Economics*. Edited by Bernt P. Stigum. Princeton: Princeton University Press, pp. 354–78.

Campos, Julia, David F. Hendry, and Hans-Martin Krolzig. 2003. Consistent model selection by an automatic *Gets* approach. *Oxford Bulletin of Economics and Statistics* 65: 803–19. [CrossRef]

Castle, Jennifer L., Jurgen A. Doornik, and David F. Hendry. 2012. Model selection when there are multiple breaks. *Journal of Econometrics* 169: 239–46. [CrossRef]

Castle, Jennifer L., Jurgen A. Doornik, and David F. Hendry. 2021. Forecasting principles from experience with forecasting competitions. *Forecasting* 3: 138–65. [CrossRef]

Castle, Jennifer L., Jurgen A. Doornik, David F. Hendry, and Felix Pretis. 2015. Detecting location shifts during model selection by step-indicator saturation. *Econometrics* 3: 240–64. [CrossRef]

Castle, Jennifer L., Michael P. Clements, and David F. Hendry. 2015. Robust approaches to forecasting. *International Journal of Forecasting* 31: 99–112. [CrossRef]

Chu, Chia-Shang, Maxwell Stinchcombe, and Halbert White. 1996. Monitoring structural change. *Econometrica* 64: 1045–65. [CrossRef]

Clements, Michael P., and David F. Hendry. 1993. On the limitations of comparing mean squared forecast errors (with discussion). *Journal of Forecasting* 12: 617–37. Reprinted in Mills, Terence C., ed. 1999. *Economic Forecasting*. Cheltenham: Edward Elgar Publishing.

Clements, Michael P., and David F. Hendry. 1998. *Forecasting Economic Time Series*. Cambridge: Cambridge University Press.

Clements, Michael P., and David F. Hendry. 2001. Explaining the results of the M3 forecasting competition. *International Journal of Forecasting* 17: 550–54.

Doornik, Jurgen A. 2009. Autometrics. In *The Methodology and Practice of Econometrics: A Festschrift in Honour of David F. Hendry*. Edited by Jennifer L. Castle and Neil Shephard. Oxford: Oxford University Press, pp. 88–121.

Doornik, Jurgen A. 2018. *Object-Oriented Matrix Programming Using Ox*, 8th ed. London: Timberlake Consultants Press.

Doornik, Jurgen A., Jennifer L. Castle, and David F. Hendry. 2020a. Card forecasts for M4. *International Journal of Forecasting* 36: 129–34. [CrossRef]

Doornik, Jurgen A., Jennifer L. Castle, and David F. Hendry. 2020b. Short-term forecasting of the coronavirus pandemic. *International Journal of Forecasting*, in press. [CrossRef] [PubMed]

Fildes, Robert, and Keith Ord. 2002. Forecasting competitions—Their role in improving forecasting practice and research. In *A Companion to Economic Forecasting*. Edited by Michael P. Clements and David F. Hendry. Oxford: Blackwells, pp. 322–53.

Hendry, David F. 2006. Robustifying forecasts from equilibrium-correction models. *Journal of Econometrics* 135: 399–426. [CrossRef]

Hendry, David F., and Grayham E. Mizon. 2012. Open-model forecast-error taxonomies. In *Recent Advances and Future Directions in Causality, Prediction, and Specification Analysis*. Edited by Xiaohong Chen and Norman R. Swanson. New York: Springer, pp. 219–40.

Hendry, David F., and Jurgen A. Doornik. 2018. *Empirical Econometric Modelling—PcGive 15 Volume I*. London: Timberlake Consultants Press.

Ing, Ching-Kang, and Ching-Zong Wei. 2003. On same-realization prediction in an infinite-order autoregressive process. *Journal of Multivariate Analysis* 85: 130–55. [CrossRef]

Leeb, Hannes, and Benedikt M. Pötscher. 2009. Model selection. In *Handbook of Financial Time Series*. Edited by Torben Andersen, Richard A. Davis, Jens-Peter Kreiss and Thomas Mikosch. Berlin: Springer, pp. 889–926.

Makridakis, Spyros, and Michele Hibon. 2000. The M3-competition: Results, conclusions and implications. *International Journal of Forecasting* 16: 451–76. [CrossRef]

Makridakis, Spyros, Evangelos Spiliotis, and Vassilios Assimakopoulos. 2020. The M4 competition: 100,000 time series and 61 forecasting methods. *International Journal of Forecasting* 36: 54–74. [CrossRef]

Pötscher, Benedikt M. 1991. Effects of model selection on inference. *Econometric Theory* 7: 163–85. [CrossRef]

Shibata, Ritei. 1980. Asymptotically efficient selection of the order of the model for estimating parameters of a linear process. *Annals of Statistics* 8: 147–64. [CrossRef]

Stock, James, and Mark W. Watson. 2009. Phillips curve inflation forecasts. In *Understanding Inflation and the Implications for Monetary Policy*. Edited by Jeff Fuhrer, Yolanda Kodrzycki, Jane Sneddon Little and Giovanni Olivei. Cambridge: MIT Press, pp. 99–202.