ISSN 1471-0498

DEPARTMENT OF ECONOMICS

DISCUSSION PAPER SERIES

Policy Analysis, Forediction, and Forecast Failure

Jennifer L. Castle, David F. Hendry and Andrew B. Martinez

Number 809
October, 2016

Manor Road Building, Oxford OX1 3UQ

Policy Analysis, Forediction, and Forecast Failure

Jennifer L. Castle, David F. Hendry and Andrew B. Martinez
Department of Economics, and Institute for New Economic Thinking,
Oxford Martin School, University of Oxford, UK.*

October 18, 2016

Abstract

Economic policy agencies accompany forecasts with narratives, a combination we call *foredictions*, often basing policy changes on developments envisaged. Forecast failure need not impugn a forecasting model, although it may, but almost inevitably entails forediction failure and invalidity of the associated policy. Most policy regime changes involve location shifts, which can induce forediction failure unless the policy variable is super exogenous in the policy model. We propose a step-indicator saturation test to check in advance for invariance to policy changes. Systematic forecast failure, or a lack of invariance, previously justified by narratives reveals such stories to be economic fiction.

Keywords: Forediction; Invariance; Super exogeneity; Indicator saturation; Co-breaking; *Autometrics*.
JEL classifications: C51, C22.

1 Introduction

The term 'forediction' was coined in Hendry (2001) as a neologism to replace 'forecasting', because such a term had not yet acquired the many negative connotations attached to words of greater antiquity. Prediction (in science) and precast (of concrete) already had well-defined usages, leaving just forediction.[1] As Hendry composed at the time, using 11 of the more than 30 synonyms for 'looking into the future' that he discussed:

> "Those who can, do; those who can't, *forecast*" spoke the *foresightful oracle* of *Delphi* when she *divined* the future to *foretell* the *prophecy* by a *soothsayer* whose *premonitions* included the *expectation* that one day economists would be able to *predict* the impact of economic policy....

Many of these older synonyms now have adverse connotations because their origins lie in methods that are now discredited. The prevalence of systematic forecast failure during the 'Great Recession' has further eroded the less-than-high reputation of economic forecasting, about which jokes already abounded, so the word 'forecasting' is in danger of joining that group.

More importantly, forediction also captures the notion of a close link between a forecast and an accompanying narrative, so is intended to convey a **fore**cast made alongside a story (**diction**) describing

*This research was supported in part by grants from the Robertson Foundation, Institute for New Economic Thinking and Statistics Norway, Project 236935/H20. We are indebted to Jurgen A. Doornik, Neil R. Ericsson, Bent Nielsen, Ragnar Nymoen and Felix Pretis for helpful comments on an earlier version. Contact details: jennifer.castle@magd.ox.ac.uk; david.hendry@nuffield.ox.ac.uk; andrew.martinez@economics.ox.ac.uk.

[1] Archival search has since revealed two earlier, somewhat unrelated, uses of forediction: (1) by Richard Feynman in a letter to Enrico Fermi in 1951 (see CBPF-CS-012/97); and (2) in Stenner (1964).

that forecast verbally: as Leamer (2009) remarks, 'humans are pattern-seeking story-telling animals'. That narrative is also often used to justify a policy decision that is claimed to be entailed by the forecast. Links between forecasts and their accompanying narratives have been brought into new salience through innovative recent papers by Stekler and Symington (2016) and Ericsson (2016b). The former develop quantitative indexes of optimism/pessimism about the US economy by analyzing the qualitative information in the minutes from Federal Open Market Committee (FOMC) meetings, scaled to match real GDP growth rates in percentages per annum. The latter calibrates the Stekler and Symington (2016) index for real GDP, denoted FMI, to past GDP outcomes, removing its truncation to $(-1, +4)$, and shows that the resulting index can provide excellent post-casts of the Fed's 'Greenbook' forecasts of 2006–10 (which are only released after a 5-year delay). Clements and Reade (2016) consider whether Bank of England narratives provide additional information beyond their inflation forecasts.

We call a published numerical forecast a 'direct forecast', and one consructed from the narrative a 'derived forecast'. Finding a close link between a direct forecast and the associated derived forecast constructed from the accompanying narrative entails that although systematic forecast failure need not reject the empirical econometric system used for forecasting (see e.g., Clements and Hendry, 1999), nevertheless failure of the derived forecast reveals the related narrative to be fictional, which we call forediction failure. We seek to establish when systematic forecast failure entails forediction failure and when that in turn implies policy invalidity. We also investigate whether failures of invariance in policy models are detectable in advance, which should enable more robust policy models to be developed.

There are four key steps in the process of evaluating foredictions and associated policies. The first is to establish whether the direct and derived forecasts are almost the same. Then forecasts are indeed closely described by the accompanying narrative in the sense formalized by Ericsson (2016b), namely accurate estimates of the forecasts can be derived by quantifying the narrative. For example, regressing the derived forecasts on the direct should deliver a highly significant relation. If so, the second step follows when systematic failure of the direct forecasts occurs, for which many tests exist. This second step is the essence of forediction failure, checked by testing (a) if the narrative-derived forecasts remain linked to the direct forecasts, and (b) they also differ significantly from the realized outcomes. In that case, the accompanying narrative must also be rejected. Thirdly, if a policy implementation had been justified by the forediction, then it too must be invalid after forediction failure. This is harder to quantify unless there is a known policy rule, such as a Taylor rule linking interest rate responses to inflation, but the policy link may be stated explicitly in the narrative. The fourth step is to test if the forediction failure actually resulted from the policy change itself because of a lack of invariance of the model used in the policy analysis to that change. In that case, the policy model is also shown to be invalid.

Evaluating the validity of policy analysis involves two intermediate links. First, genuine causality from policy instruments to target variables is obviously essential for policy effectiveness if changing the policy variable is to affect the target (see e.g., Cartwright, 1989). That influence could be indirect, as when an interest rate affects aggregate demand which thereby changes the rate of inflation. Secondly, the invariance of the parameters of the empirical models used for policy to shifts in the distributions of their explanatory variables is also essential for analyses to correctly represent the likely outcomes of policy changes. Thus, a policy model must not only embody genuine causal links in the data generation process (DGP), invariance also requires the absence of links between target parameters and policy-instrument parameters, since the former cannot be invariant to changes in the latter if their DGP parameters are linked, independently of what modellers may assume: see e.g., Hendry (2004) and Zhang, Zhang, and Schölkopf (2015). Consequently, weak exogeneity matters in a policy context because its two key requirements are that the relevant model captures the parameters of interest for policy analyses, and that the parameters of the policy instrument and the target processes are unconnected in the DGP, not merely in their assumed parameter spaces. Since changes in policy are involved, valid policy then requires super exogeneity of the policy variables in the target processes: see Engle, Hendry, and Richard (1983).

Combining these ideas, forediction evaluation becomes a feasible approach to checking the validity of policy decisions based on forecasts linked to a narrative that the policy agency wishes to convey. Policy decisions can fail because empirical models are inadequate, causal links do not exist, parameters are not invariant, the story is incorrect, or unanticipated events occur. By checking the properties of forecasts and the invariance of the policy model to the policy change envisaged, the source of forediction failure may be more clearly discerned, hopefully leading to fewer bad mistakes in the future.

The structure of the paper is as follows. Section 2 explores forediction as linking forecasts with accompanying narratives that are often used to justify policy decisions, where §2.1 discusses whether narratives or forecasts come first, or are jointly decided. Section 3 considers what can be learned from systematic forecast failure, and whether there are any entailed implications for forediction failure. More formally, §3.1 outlines a simple theoretical policy DGP; §3.2 describes what can be learned from systematic forecast failure using a taxonomy of forecast errors for a mis-specified model of that DGP; §3.3 provides a numerical illustration; and §3.4 notes some of the implications of forediction failure for intertemporal economic theories. Section 4 presents a two-stage test for invariance using automatic model selection based step-indicator saturation (SIS: see Castle, Doornik, Hendry, and Pretis, 2015), which extends impulse-indicator saturation (IIS: see Hendry, Johansen, and Santos, 2008). Section 5 reports some simulation evidence on its first-stage null retention frequency of irrelevant indicators (gauge) and its retention frequency of relevant indicators (potency), as well as its second-stage rejection frequencies in the presence and absence of invariance. Section 6 applies this SIS-based test to the small artificial-data policy model analyzed in §3.1, and §6.1 investigates whether a shift in a policy parameter can be detected using multiplicative indicator saturation (MIS), where step indicators are interacted with variables. Section 7 considers how intercept corrections can improve forecasts without changing the forecasting model's policy responses. Section 8 concludes.

2 Forediction: linking forecasts, narratives, and policies

For many years, Central Banks have published narratives both to describe and interpret their forecasts, and often justify entailed policies. Important examples include the minutes from Federal Open Market Committee (FOMC) meetings, the Inflation Reports of the Bank of England, and International Monetary Fund (IMF) Reports. For Banca d'Italia, Siviero and Terlizzese (2001) state that:

> ...forecasting does not simply amount to producing a set of figures: rather, it aims at assembling a fully-fledged view—one may call it a "story behind the figures"—of what could happen: a story that has to be internally consistent, whose logical plausibility can be assessed, whose structure is sufficiently articulated to allow one to make a systematic comparison with the wealth of information that accumulates as time goes by.

Such an approach is what we term forediction, and any claims as to its success need to be evaluated in the light of the widespread forecast failure precipitated by the 2008–9 Financial Crisis. As we show below, closely tying narratives and forecasts may actually achieve the opposite of what those authors seem to infer, namely end by rejecting both the narratives and associated policies when forecasts go wrong.

Sufficiently close links between direct forecasts and forecasts derived from their accompanying narratives, as highlighted by Stekler and Symington (2016) and Ericsson (2016b), entail that systematic forecast failure vitiates any related narrative and its associated policy. This holds irrespective of whether the direct forecasts led to the narrative, or the forecasts were modified (e.g., by 'judgemental adjustments') to satify a preconceived view of the future expressed in a narrative deliberately designed to justify a policy, or the two were iteratively adjusted as in the above quote, perhaps to take account of informal information available to a panel such as the Bank of England's Monetary Policy Committee (MPC). In all three cases, if the direct and derived forecasts were almost the same, and the narratives reflected cognitive models or theories, then the large forecast errors around the 'Great Recession' would also refute such thinking as addressed in Section 3. However, when direct and derived forecasts are not tightly linked, failure in one need not entail failure in the other, but both can be tested directly.

2.1 Do narratives or forecasts come first?

Forcing internal consistency between forecasts and narratives, as both Siviero and Terlizzese (2001) and Pagan (2003) stress, could be achieved by deciding the story, then choosing add factors to achieve it, or *vice versa*, or a combination of adjustments. The former authors appear to suggest the third was common at Banca d'Italia, and the fact that Bank of England forecasts are those of the MPC based on the information it has, which includes forecasts from the Bank's models, suggests mutual adjustments of forecasts and narratives, as in e.g., Bank of England (2015, p.33) (our italics):

> The projections for growth and inflation are underpinned by four key *judgements*.

Not only could add factors be used to match forecasts to a narrative, if policy makers had a suite of forecasting models at their disposal, then the weightings on different models could be adjusted to match their pooled forecast to the narrative, or both could be modified in an iterative process. Genberg and Martinez (2014) show the link between narratives and forecasts at the International Monetary Fund (IMF), where forecasts are generated on a continuous basis through the use of a spreadsheet framework that is occasionally supplemented by satellite models, as described in Independent Evaluation Office (2014). Such forecasts 'form the basis of the analysis [...] and of the [IMF's] view of the outlook for the world economy' (p.1). Thus, adjustments to forecasts and changes to the associated narratives tend to go hand-in-hand at many major institutions.

In a setting with several forecasting models used by an agency that nevertheless delivered a unique policy prescription, possibly based on a 'pooled forecast', then to avoid implementing policy incorrectly, super exogeneity with respect to the policy instrument would seem essential for every model used in that forecast. In practice, the above quotes suggest some Central Banks act as though there is a unique policy model constrained to match their narrative, where all the models in the forecasting suite used in the policy decision are assumed to be invariant to any resulting changes in policy. This requirement also apples to scenario studies as the parameters of the models being used must be valid across all the states examined. Simply asserting that the policy model is 'structural' because it is derived from a theory is hardly sufficient. Akram and Nymoen (2009) consider the role of empirical validity in monetary policy, and caution against over-riding it. Even though policy makers recognise that their policy model may not be the best forecasting model, as illustrated by the claimed trade-off between theory consistency and empirical coherence in Pagan (2003), forecast failure can still entail forediction failure, as shown in Section 3.

2.2 Is there a link between forecasts and policy?

Evaluating policy-based forediction failure also requires there to be a link between the forecasts and policy changes. This can either occur indirectly through the narratives or directly through the forecasts themselves (e.g. through a forward looking Taylor rule). That direct forecasts and forecasts derived from the narratives are closely linked implies that policies justified by those narratives, are also linked to the forecasts. However, is there any evidence of such links between forecasts and policy?

In their analysis of the differences between the Greenbook and FOMC forecasts, Romer and Romer (2008) find that there is a statistically significantly link between differences in these two forecasts and monetary policy shocks. This can be interpreted as evidence suggesting that policy-makers forecasts influence their decisions above and beyond other informational inputs. As such, the policy decision is both influenced by and possibly influencing the forecasts.

Regardless of this influence, the link between forecasts and policy depends on how the forecasts are used. Ellison and Sargent (2012) illustrate that it is possible for 'bad forecasters' to be 'good policy-makers'. In this sense forecast failure could potentially be associated with policy success. However, this is less likely when focusing on systematic forecast failure. Sinclair, Tien, and Gamber (2016) find that forecast failure at the Fed is associated with policy failure. They argue that in 2007–08 the Greenbook's overpredictions of output and inflation "resulted in a large contractionary [monetary policy] shock" of around 7 percentage points. By their calculations, using a forward-looking Taylor rule, this forecast-error-induced shock reduced real growth by almost 1.5 percentage points. Furthermore, Independent Evaluation Office (2014) illustrates that in IMF Programs, the forecasts can effectively be interpreted as the negotiated policy targets. This suggests that forecasts, narratives and policy are closely linked so that systematic forediction failure can be used to refute the validity of policy decisions.

3 Forecast failure and forediction failure

The important distinction between the implications of systematic forecast failure and of forediction failure can be illustrated by the oxygen cylinder explosion on Apollo 13. That unfortunate event led to a massive forecast failure of its arrival time on the moon, but even so, most certainly did not reject either Newton's laws or NASA's forecasting algorithms. Indeed, NASA used both that gravitational theory and its forecasting algorithms to correctly track the subsequent motion of the spacecrafts module and bring the astronauts safely back to Earth. Consequently, forecast failure need not reject the forecasting model or the underlying theory from which that model was derived (also see Castle and Hendry, 2011). However, the plight of Apollo 13 did refute NASA's earlier narrative of its time to landing on the moon, and also refuted the policy claims in that narrative of what the mission would achieve, so forediction failure and policy invalidity occurred.

Concerning the rejection of Keynesian models in the 1970s, Hendry and Mizon (2011b) argue:
"This arose in part because the poor forecasting record of empirically-based macro-econometric models led to the non-sequitur that their underlying theory was incorrect, on a par with rejecting Newton's laws of universal gravitation because the forecasts of Apollo 13's arrival on the moon went horrendously wrong."
Thus, systematic forecast failure **by itself** is not a sufficient condition for rejecting an underlying theory or any asociated forecasting model, and conversely, forecasting success may, but need not, 'corroborate' the forecasting model and its supporting theory since robust forecasting devices can outperform the forecasts from an estimated in-sample DGP after a location shift (see Hendry, 2006).

Nevertheless, when several rival explanations exist, forecast failure can play an important role in distinguishing between them as discussed by Spanos (2007). For example, the failure of clocks on rapidly orbiting satellites to match earth-bound clocks refutes Newtonian theory due to neglecting relativistic effects, so here failure of the forecast that 'the clocks will still match afer a decade' rejects the theory. Moreover, systematic forecast failure seems to almost always reject any narrative associated with the failed forecasts and any policy implications therefrom, so inevitably results in forediction failure.

3.1 A simple policy data generation process

In this subsection, we develop a simple theoretical policy model which is mis-specified for its economy's DGP. At time T, the DGP shifts unexpectedly, so there is forecast failure. However, even if the DGP did not shift, because the policy model is mis-specified a policy change itself can induce forecast failure. In our model, the policy agency decides on an action at precisely time T, so the two shifts interact. We provide a taxonomy of the resulting sources of failure, and what can be inferred from each component.

Let y_{t+1} be a policy target, say inflation, z_t the instrument an agency controls to influence that target, say interest rates, where x_{t+1} is the variable that is directly influenced by the instrument, say aggregate excess demand, which directly affects inflation. Also, let $w_{1,t+1}$ represent the net effect of other variables on the target, say world inflation in the domestic currency, and $w_{2,t+1}$ denote additional forces directly affecting domestic demand, where the $\{w_{i,t+1}\}$ are super exogenous. The system is formalized as:

$$y_t = \gamma_0 + \gamma_1 x_t + \gamma_2 w_{1,t} + \epsilon_t \tag{1}$$

where for simplicity $\epsilon_t \sim \mathsf{IN}[0, \sigma_\epsilon^2]$ with $\gamma_1 > 0$ and $\gamma_2 > 0$; and :

$$x_t = \beta_0 + \beta_1 z_{t-1} + \beta_2 w_{2,t} + \nu_t \tag{2}$$

with $\nu_t \sim \mathsf{IN}[0, \sigma_\nu^2]$ where $\beta_1 < 0$ and $\beta_2 > 0$. In the next period, solving as a function of policy and exogenous variables:

$$y_{t+1} = (\gamma_0 + \gamma_1\beta_0) + \gamma_1\beta_1 z_t + \gamma_2 w_{1,t+1} + \gamma_1\beta_2 w_{2,t+1} + (\epsilon_{t+1} + \gamma_1\nu_{t+1}) \tag{3}$$

Consequently, a policy shift moving z_t to $z_t + \delta$ would, on average and *ceteris paribus*, move y_{t+1} to $y_{t+1} + \delta\gamma_1\beta_1$. We omit endogenous dynamics to focus on the issues around forediction.

In a high-dimensional, wide-sense non-stationary world, where the model is not the DGP, and parameters need to be estimated from data that are not accurately measured, the policy agency's model of inflation is the mis-specified representation:

$$y_t = \lambda_0 + \lambda_1 x_t + e_t \tag{4}$$

and its model of demand is:

$$x_t = \theta_0 + \theta_1 z_{t-1} + u_t \tag{5}$$

This system leads to its view of the policy impact of z_t at $t + 1$ as being on average and *ceteris paribus*:

$$y_{t+1|t} = (\lambda_0 + \lambda_1\theta_0) + \lambda_1\theta_1 z_t + (e_{t+1} + \lambda_1 u_{t+1}) \tag{6}$$

so a policy shift moving z_t to $z_t + \delta$ would be expected to move $y_{t+1|t}$ to $y_{t+1|t} + \delta\widehat{\lambda}_1\widehat{\theta}_1$ where $\widehat{\lambda}_1$ and $\widehat{\theta}_1$ are in-sample estimates of λ_1 and θ_1 respectively. If the agency wishes to lower inflation when $\lambda_1 > 0$ and $\theta_1 < 0$, it must set $\delta > 0$ such that $\delta\widehat{\lambda}_1\widehat{\theta}_1 < 0$ (e.g., -0.01 corresponds to a 1 percentage point reduction in the annual inflation rate).

The equilibrium means in the stationary world before any shifts or policy changes are:

$$\mathsf{E}[y_t] = \gamma_0 + \gamma_1\beta_0 + \gamma_1\beta_1\mu_z + \gamma_2\mu_{w_1} + \gamma_1\beta_2\mu_{w_2} = \mu_y$$

where $\mathsf{E}[z_t] = \mu_z$, $\mathsf{E}[w_{1,t}] = \mu_{w_1}$ and $\mathsf{E}[w_{2,t}] = \mu_{w_2}$, so that $\lambda_0 + \lambda_1\mu_x = \mu_y$ as well from (4), and:

$$\mathsf{E}[x_t] = \beta_0 + \beta_1\mu_z + \beta_2\mu_{w_2} = \mu_x$$

where from (5), $\theta_0 + \theta_1\mu_z = \mu_x$. We now consider the effects of the unexpected shift in the DGP in §3.1.1, of the policy change in §3.1.2, and of mis-specification of the policy model for the DGP in §3.1.3.

3.1.1 The DGP shifts unexpectedly

Just as the policy change is implemented at time T, the DGP shifts unexpectedly to:

$$y_{T+1} = \gamma_0^* + \gamma_1^* x_{T+1} + \gamma_2^* w_{1,T+1} + \epsilon_{t+1} \tag{7}$$

and:

$$x_{T+1} = \beta_0^* + \beta_1^* z_T + \beta_2^* w_{2,T+1} + \nu_{T+1} \tag{8}$$

For simplicity we assume the error distributions remain the same, so that:

$$y_{T+1} = (\gamma_0^* + \gamma_1^*\beta_0^*) + \gamma_1^*\beta_1^* z_T + \gamma_2^* w_{1,T+1} + \gamma_1^*\beta_2^* w_{2,T+1} + (\epsilon_{T+1} + \gamma_1^*\nu_{T+1}) \tag{9}$$

The equilibrium means after the shift and before the policy change are:

$$\mathsf{E}_{T+1}\left[y_{T+1|T}\right] = \mu_{y^*} = \gamma_0^* + \gamma_1^*\beta_0^* + \gamma_1^*\beta_1^*\mu_z + \gamma_2^*\mu_{w_1} + \gamma_1^*\beta_2^*\mu_{w_2}$$

where the expectation is taken at time $T+1$, and when $\mu_z, \mu_{w_1}, \mu_{w_2}$ remain unchanged:

$$\mu_{x^*} = \beta_0^* + \beta_1^*\mu_z + \beta_2^*\mu_{w_2}$$

Even using the in-sample DGP (3) with known parameters and known values for the exogenous variables, there will be forecast failure for a large location shift from μ_y to μ_y^* since for $\epsilon_{T+1|T} = y_{T+1} - y_{T+1|T}$:

$$\begin{aligned}\epsilon_{T+1|T} = \mu_{y^*} - \mu_y &+ (\gamma_1^*\beta_1^* - \gamma_1\beta_1)(z_T - \mu_z) + (\gamma_2^* - \gamma_2)\left(w_{1,T+1} - \mu_{w_1}\right)\\ &+ (\gamma_1^*\beta_2^* - \gamma_1\beta_2)\left(w_{2,T+1} - \mu_{w_2}\right) + (\epsilon_{T+1} + \gamma_1^*\nu_{T+1})\end{aligned}$$

so that:

$$\mathsf{E}_{T+1}\left[\epsilon_{T+1|T}\right] = \mu_{y^*} - \mu_y \tag{10}$$

and similarly for $\nu_{T+1|T} = x_{T+1} - x_{T+1|T}$, then $\mathsf{E}_{T+1}\left[\nu_{T+1|T}\right] = \mu_{x^*} - \mu_x$.

3.1.2 The policy change

Next, the policy shift changes z_T to $z_T + \delta$, which is a mistake as that policy leads to the impact $\delta\gamma_1^*\beta_1^*$ instead of the anticipated $\delta\gamma_1\beta_1$ when the agency knew the in-sample DGP. Thus, the policy shift would alter the mean forecast error to:

$$\mathsf{E}_{T+1}\left[\epsilon_{T+1|T} \mid \delta \neq 0\right] = \mu_{y^*} - \mu_y + (\gamma_1^*\beta_1^* - \gamma_1\beta_1)\,\delta \tag{11}$$

The additional component in (11) compared to (10) would be zero if $\delta = 0$, so the failure of super exogeneity of the policy variable for the target augments, or depending on signs, possibly attenuates the forecast failure. Importantly, if $\mu_{y^*} = \mu_y$ the policy shift by itself could create forecast failure. Hence section 4 addresses testing for such a failure in advance of a policy-regime shift.

3.1.3 The role of mis-specification

Third, there would be a mis-specification effect from using (6), as the scenario calculated *ex ante* would suggest the effect to be $\delta\lambda_1\theta_1$. Although there is also an estimation effect, it is probably $\mathsf{O}_p(T^{-1})$, but to complete the taxonomy of forecast errors below, we will add estimation uncertainty since the parameters of (6) could not be 'known'. Denoting such estimates of the model by $\widetilde{\ }$, then:

$$\mathsf{E}\left[y_t\right] = \mu_y = \mathsf{E}\left[(\widetilde{\lambda}_0 + \widetilde{\lambda}_1\widetilde{\theta}_0) + \widetilde{\lambda}_1\widetilde{\theta}_1 z_{t-1}\right] \approx (\lambda_0 + \lambda_1\theta_0) + \lambda_1\theta_1\mu_z$$

where both z_T and μ_z are almost bound to be known to the policy agency. From the $T+1$ DGP in (9):

$$y_{T+1} = \mu_y^* + \gamma_1^*\beta_1^*(z_T - \mu_z) + \gamma_2^*\left(w_{1,T+1} - \mu_{w_1}\right) + \gamma_1^*\beta_2^*\left(w_{2,T+1} - \mu_{w_2}\right) + (\epsilon_{T+1} + \gamma_1^*\nu_{T+1}) \tag{12}$$

but from (6), the agency would forecast y_{T+1} as:

$$\widetilde{y}_{T+1|T} = (\widetilde{\lambda}_0 + \widetilde{\lambda}_1\widetilde{\theta}_0) + \widetilde{\lambda}_1\widetilde{\theta}_1\widetilde{z}_T = \widetilde{\mu}_y + \widetilde{\lambda}_1\widetilde{\theta}_1 (\widetilde{z}_T - \mu_z) \tag{13}$$

where $\widetilde{z}_T$ is the measured forecast-origin value. An agency is almost certain to know the correct value, but we allow for the possibility of an incorrect forecast-origin value to complete the taxonomy of forecast-failure outcomes in (15) and Table 1. Then the agency anticipates that by shifting z_T to $z_T + \delta$ would on average revise the outcome to $y_{T+1|T} + \delta\widetilde{\lambda}_1\widetilde{\theta}_1$, as against the actual outcome in (12), leading to a forecast error of $\widetilde{v}_{T+1|T} = y_{T+1} - \widetilde{y}_{T+1|T}$:

$$\begin{aligned}\widetilde{v}_{T+1|T} = {} & \left(\mu_y^* - \widetilde{\mu}_y\right) + \left(\gamma_1^*\beta_1^* - \widetilde{\lambda}_1\widetilde{\theta}_1\right)(z_T - \mu_z + \delta) \\ & + \gamma_2^*\left(w_{1,T+1} - \mu_{w_1}\right) + \gamma_1^*\beta_2^*\left(w_{2,T+1} - \mu_{w_2}\right) + \left(\epsilon_{T+1} + \gamma_1^*\nu_{T+1}\right)\end{aligned} \tag{14}$$

Unlike NASA, unless the model is revised, the same average error will occur in the following periods leading to systematic forecast failure. Indeed, even if the policy agency included $w_{1,t+1}$ and $w_{2,t+1}$ appropriately in their forecasting model's equations, their roles in (14) would be replaced by any shifts in their parameter values (see Hendry and Mizon, 2012).

3.1.4 The sources of forecast failure

The error in (14) can be decomposed into terms representing mis-estimation (labelled *(a)*), mis-specification (*(b)*) and change (*(c)*) for each of the equilibrium mean (labelled *(i)*), slope parameter (*(ii)*) and unobserved terms (*(iii)*), as in (15). The implications of this 3x3 framework are recorded in the Taxonomy Table 1 under the assumption that the derived forecasts are closely similar to the direct.

$$\begin{aligned}\widetilde{v}_{T+1|T} \simeq {} & \left(\mu_{y,p} - \widetilde{\mu}_y\right) && i(a) \\ & + \left(\mu_y - \mu_{y,p}\right) && i(b) \\ & + \left(\mu_y^* - \mu_y\right) && i(c) \\ & + \left(\lambda_1\theta_1 - \widetilde{\lambda}_1\widetilde{\theta}_1\right)(z_T - \mu_z + \delta) && ii(a) \\ & + (\gamma_1\beta_1 - \lambda_1\theta_1)(z_T - \mu_z + \delta) && ii(b) \\ & + (\gamma_1^*\beta_1^* - \gamma_1\beta_1)(z_T - \mu_z + \delta) && ii(c) \\ & + \lambda_1\theta_1 (z_T - \widetilde{z}_T) && iii(a) \\ & + \gamma_2^*\left(w_{1,T+1} - \mu_{w_1}\right) + \gamma_1^*\beta_2^*\left(w_{2,T+1} - \mu_{w_2}\right) && iii(b) \\ & + \left(\epsilon_{T+1} + \gamma_1^*\nu_{T+1}\right) && iii(c)\end{aligned} \tag{15}$$

In (15), the impacts of each of the nine terms can be considered in isolation by setting all the others to zero in turn. The plims of estimates are denoted by an additional subscript $_p$ to focus on location shift effects and policy errors in the taxonomy discussed in the next section. Although only *i(c)* is likely to cause forecast failure, any narrative and associated policy as discussed above may well be rejected by systematic errors arising from any of these nine terms.

3.2 What can be learned from systematic forecast failure?

The three possible mistakes—mis-estimation, mis-specification and change—potentially affect each of the three main components of any forecasting model—equilibrium means, slope parameters and unobserved terms—lead to the nine terms in Table 1. This condensed model-free taxonomy of sources of forecast errors is based on Ericsson (2016a), which built on those in Clements and Hendry (1998) for closed models and Hendry and Mizon (2012) for open systems.

As not all sources of systematic forecast failure lead to the same implications, we consider which sources should cause rejection of the forecasting model (denoted FM in the table), the forediction narrative (FN), or impugn policy validity (PV). Systematic failure of the forecasts could arise from any of the terms, but the consequences vary with the source of the mistake. We assume that the derived forecasts are sufficiently close to the direct that both suffer systematic forecast failure, so in almost every case in the table, the entailed forediction and policy are rejected. However, a '?' denotes that systematic forecast failure is less likely in the given case, but if it does occur then the consequences are as shown.

Component	Source		
	Mis-estimation	Mis-specification	Change
Equilibrium mean	*i*(*a*)	*i*(*b*)	*i*(*c*)
Reject	FN?, PV?	FM, FN, PV	FM, FN, PV
Slope parameter	*ii*(*a*)	*ii*(*b*)	*ii*(*c*)
Reject (if $\delta \neq 0$)	FN?, PV?	FM, FN, PV	FM, FN, PV
Unobserved terms	*iii*(*a*) [forecast origin]	*iii*(*b*) [omitted variable]	*iii*(*c*) [innovation error]
Reject	FN, PV	FM?, FN?, PV?	FN?, PV?

Table 1: A taxonomy of the implications of systematic forecast failures.

Taking the columns in turn, mis-estimation by itself will rarely lead to systematic forecast failure and should not lead to rejecting the **formulation** of a forecasting model, or the theory from which it was derived, as appropriate estimation would avoid such problems. For example,when the forecast origin has been mis-measured by a statistical agency resulting in a large forecast error in *iii(a)*, one should not reject either the forecasting model or the underlying theory. Nevertheless, following systematic forecast failure, any policy conclusions that had been drawn and their accompanying narrative should be rejected as incorrect, at least until more accurate data are produced.

Next, mis-specification generally entails that the model should be rejected against a less badly specified alternative, and that foredictions and policy claims must be carefully evaluated as they too will often be invalid. In stationary processes, mis-specification alone will not lead to systematic forecast failure, but the real economic world is not stationary, and policy regime shifts will, and are intended to, alter the distributions of variables. That omitted variables need not induce systematic forecast failure may surprise at first sight, but *iii(b)* in (15) reveals their *direct* effect has a zero mean as shown by Hendry and Mizon (2012).

Third, unanticipated changes in the non-zero-mean components of forecasting models will usually cause systematic forecast failure. However, slope shifts on zero-mean influences and changes in innovation error variances need not. Nevertheless, in almost all cases systematic forecast failure will induce both forediction and policy failure if the policy implemented from the mistaken narrative depended on a component that had changed. Finally, the consequences for any underlying theory depend on how that theory is formulated, an issue briefly discussed in §3.4. Table 2 relates the taxonomy in Table 1 to the sources of the forecast errors from (15) to illustrate which indicator-saturation test could be used, where the order (SIS, IIS) etc. shows their likely potency. SIS^{se} denotes the super exogeneity test in section 4.

When the source of forecast failure is the equilibrium mean or forecast origin mis-estimation, then SIS is most likely to detect the systematically signed forecast errors, whereas for other unobserved terms IIS is generally best equipped to detect these changes. When the slope parameter is the source of failure

Component	Problem		
	Mis-estimation	Mis-specification	Change
Equilibrium mean	$i(a)$ [uncertainty]	$i(b)$ [inconsistent]	$i(c)$ [shift]
Source	$\left(\mu_{y,p}-\widehat{\mu}_y\right)$	$+\left(\mu_y-\mu_{y,p}\right)$	$+\left(\mu_y^*-\mu_y\right)$
Test	SIS, IIS	SIS, IIS	SIS, IIS, SIS^{se}
Slope parameter	$ii(a)$ [uncertainty]	$ii(b)$ [inconsistent]	$ii(c)$ [break]
Source	$+\left(\lambda_1\theta_1-\widehat{\lambda}_1\widehat{\theta}_1\right)\times$	$+\left(\gamma_1\beta_1-\lambda_1\theta_1\right)\times$	$+\left(\gamma_1^*\beta_1^*-\gamma_1\beta_1\right)\times$
$(\delta\neq 0)$	$(z_T-\mu_z+\delta)$	$(z_T-\mu_z+\delta)$	$(z_T-\mu_z+\delta)$
Test	SIS	SIS	MIS, SIS, IIS, SIS^{se}
Unobserved terms	$iii(a)$ [forecast origin]	$iii(b)$ [omitted variable]	$iii(c)$ [innovation error]
Source	$+\lambda_1\theta_1\left(z_T-\widehat{z}_T\right)$	$+\gamma_2^*\left(w_{1,T+1}-\mu_{w_1}\right)$	$+\epsilon_{T+1}$
		$+\gamma_1^*\beta_2^*\left(w_{2,T+1}-\mu_{w_2}\right)$	$+\gamma_1^*\nu_{T+1}$
Test	SIS, IIS	IIS, SIS	IIS

Table 2: The taxonomy of systematic forecast failures with associated tests.

for $\delta\neq 0$, then SIS is generally best, whereas when $\delta=0$, IIS might help. In practice, policy invalidity and forediction failure are probably associated with *i(c)* and *ii(c)*, where both SIS and IIS tests for super exogeneity are valid. In this setting, policy failure can also be triggered through *ii(a)* and *ii(b)* which makes an SIS test for super exogeneity again attractive. Absent a policy intervention then zero-mean changes result in *ii(c)*, so may best be detected using multiplicative indicator saturation (MIS) as discussed in §6.1.

3.3 A numerical illustration

The scenario we consider is one where from (13), the agency's direct forecast, $\widetilde{y}_{T+1|T}$, is higher than the policy target value, so with $\theta_1<0$, it announces a rise in z_T of δ such that $\widetilde{y}_{T+1|T}+\delta\lambda_1\theta_1$ will be smaller than the target, and x_{T+1} will fall. Unfortunately, an unanticipated boom hits, resulting in large rises in x_{T+1} and y_{T+1}. Below, $\widehat{y}_{T+1|T}$ and $\widehat{x}_{T+1|T}$ denote forecast values from (infeasible) in-sample DGP estimates, whereas $\widetilde{y}_{T+1|T}$ and $\widetilde{x}_{T+1|T}$ are model-based forecasts, which here also correspond to the agency's forediction-based forecasts, for inflation and output. An example of how a forecast, the associated forediction and a linked policy change could be represented in this case is given in Table 3.

Forecast	Narrative	Policy
The outlook for inflation is that it will be $\mathbf{6.25}\%$ this year followed by a moderate decline to $\mathbf{5.25}\%$ next year.	Core inflation has been **elevated** in recent months. High levels of resource utilization and **high prices** of energy and other commodities have the potential to sustain **inflationary pressures** but should **moderate going forward**.	The Policy Committee decided today to raise its target interest rate by **100** basis points **due to ongoing concerns about inflation pressures**.

Table 3: Forediction example.

The numerical values of the parameters in the 5-variable in-sample simulation DGP are recorded in the top two rows of Table 4. As $w_{1,t}\sim\mathsf{IN}[\mu_{w_1},0.1]$ and $w_{2,t}\sim\mathsf{IN}[\mu_{w_2},0.1]$, their projections on x_t and z_{t-1} respectively are both zero, so $\lambda_0=6.0$, $\lambda_1=-1.0$, $\theta_0=3$, $\theta_1=1.0$, $\mu_y=6$ and $\mu_x=3$. The policy change is $\delta=1$ and the shifts in the DGP are described in bottom two rows of Table 4. Now

$\mu_{y^*} = 8.25$ and $\mu_{x^*} = 4.5$, but initially x jumps to around 6 because of the lag in its response to the policy change. For y and x these cases are denoted (*I*) & (*II*) in Figure 1, whereas cases (*III*) & (*IV*) have the same policy implementation and DGP change, but with no location shift, so $\gamma_0^* = \gamma_0 = 1.0$ and $\beta_0^* = \beta_0 = 1.0$.

in-sample	γ_0	γ_1	γ_2	σ_ϵ^2	β_0	β_1	β_2	σ_ν^2	δ	μ_z	μ_{w_1}	μ_{w_2}
value	1.0	1.0	1.0	0.1	1.0	-1.0	1.0	0.1	0	0	2	2
out-of-sample	γ_0^*	γ_1^*	γ_2^*	σ_ϵ^2	β_0^*	β_1^*	β_2^*	σ_ν^2	δ	μ_z^*	μ_{w_1}	μ_{w_2}
value	2.0	0.5	2.0	0.1	2.0	-1.5	2.0	0.1	1	1	2	2

Table 4: Parameter values for the simulation DGPs.

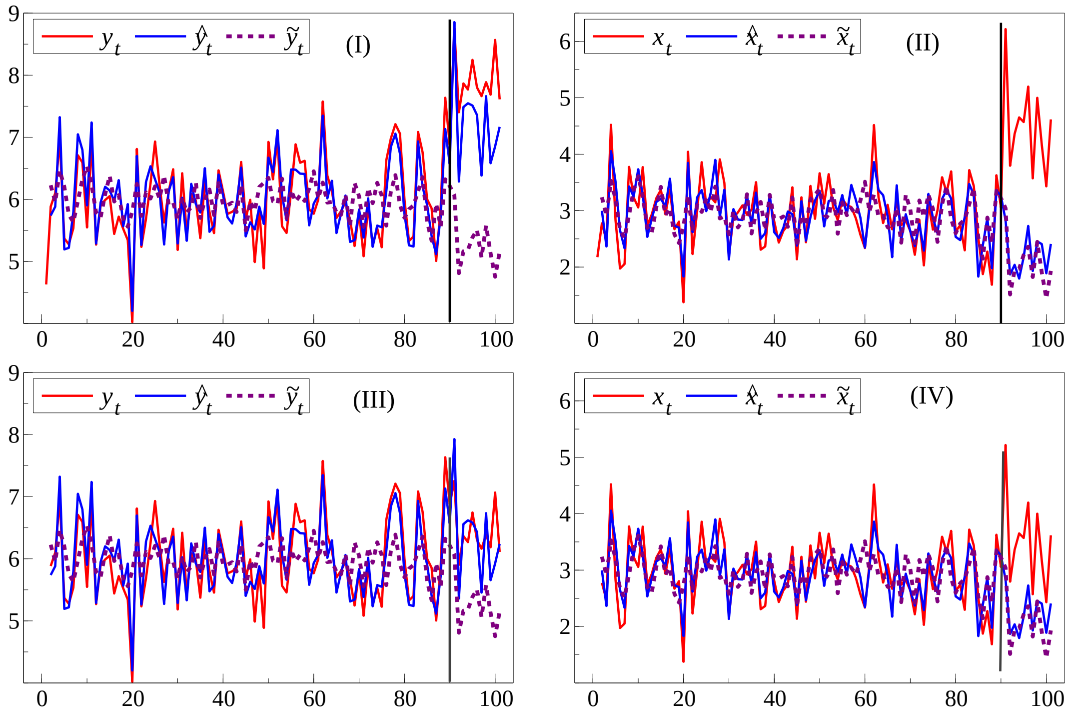

Figure 1: (I) & (II) both forediction and policy failures from the forecast failure following a changed DGP with a location shift and a policy change combined; (III) & (IV) both forediction and policy failures after a changed DGP without a location shift but with a policy change.

For a single simulation, Figure 1 (I) illustrates the systematic forecast failure of $\widehat{y}_t$ for y_t when the break in the DGP includes both a location shift and the policy change, together with the much larger failure of the model's forecast, $\widetilde{y}_t$. When there is no unexpected location shift, so $\gamma_0 = \gamma_0^*$ and $\beta_0 = \beta_0^*$, then as (III) shows, the in-sample DGP-based forecasts $\widehat{y}_t$ do not suffer failure despite the other changes in the DGP. However, the model mis-specification interacting with $\delta \neq 0$ still induces both forediction and policy failure from the model's forecast $\widetilde{y}_t$ failing. This occurs because the agency's model for y_t is mis-specified, so the δ change acts as an additional location shift, exacerbating the location shift in (I) from the change in μ_y, whereas only the policy change occurs in (III). There is little difference between $\widehat{x}_t$ and $\widetilde{x}_t$ in either scenario: even if the agency had known the in-sample DGP equation for x_t, there would have been both forediction and policy failure from the changes in the slope parameters caused by $\delta \neq 0$ as Figures 1 (II) and (IV) show.

3.4 Implications of forediction failure for inter-temporal theories

Whether or not a systematic forecast failure from whatever source impugns an underlying economic theory depends on how directly the failed model is linked to that theory. In the case of dynamic stochastic general equilibrium (DSGE) theory that link is close, so most instances of FM in Table 3.2 entail a theory failure for the DSGE class. More directly, Hendry and Mizon (2014) show that the mathematical basis of inter-temporal optimization fails when there are shifts in the distributions of the variables involved. All expectations operators must then be three-way subscripted, as in $\mathsf{E}_{\mathsf{D}_{y_t}}[y_{t+1}|\mathcal{I}_{t-1}]$ which denotes the conditional expectation of the random variable y_{t+1} formed at time t as the integral over the distribution $\mathsf{D}_{y_t}(\cdot)$ given an available information set $\mathcal{I}_{t-1}$. When $\mathsf{D}_{y_t}(\cdot)$ is the distribution $y_t \sim \mathsf{IN}[\mu_t, \sigma_y]$ where future changes in μ_t are unpredictable, then:

$$\mathsf{E}_{\mathsf{D}_{y_t}}[y_{t+1} \mid \mathcal{I}_{t-1}] = \mu_t \tag{16}$$

whereas:

$$\mathsf{E}_{\mathsf{D}_{y_{t+1}}}[y_{t+1} \mid \mathcal{I}_{t-1}] = \mu_{t+1} \tag{17}$$

so the conditional expectation $\mathsf{E}_{\mathsf{D}_{y_t}}[y_{t+1}|\mathcal{I}_{t-1}]$ formed at t is not an unbiased predictor of the outcome μ_{t+1}, and only a 'crystal-ball' predictor $\mathsf{E}_{\mathsf{D}_{y_{t+1}}}[y_{t+1}|\mathcal{I}_{t-1}]$ based on 'knowing' $\mathsf{D}_{y_{t+1}}(\cdot)$ is unbiased.

A related problem afflicts derivations which assume that the law of iterated expectations holds even when distributions shift between t and $t+1$, since then:

$$\mathsf{E}_{\mathsf{D}_{y_t}}\left[\mathsf{E}_{\mathsf{D}_{y_{t+1}}}[y_{t+1} \mid y_t]\right] \neq \mathsf{E}_{\mathsf{D}_{y_{t+1}}}[y_{t+1}] \tag{18}$$

It is unsurprising that a shift in the distribution of y_t between periods would wreck a derivation based on assuming such changes did not occur, so the collapse of DSGEs like the Bank of England's Quarterly Econometric Model (BEQEM) after the major endowment shifts occasioned by the 'Financial Crisis' should have been anticipated. A similar fate is likely to face any successor built on the same lack of mathematical foundations relevant for a changing world. Indeed, a recent analysis of the replacement model, COMPASS (Central Organising Model for Projection Analysis and Scenario Simulation) confirms that such a fate has already occurred.[2]

The results of this section re-emphasize the need to test forecasting and policy-analysis models for their invariance to policy changes prior to implementing them. Unanticipated location shifts will always be problematic, but avoiding the policy-induced failures seen in Figures 1 (III) and (IV) by testing for invariance seems feasible. Section 4 briefly describes step-indicator saturation, then explains its extension to test invariance based on including in conditional equations the significant step indicators from marginal-model analyses.

4 Step-indicator saturation (SIS) based test of invariance

Many tests for super exogeneity have been proposed since Engle, Hendry, and Richard (1983) (see e.g., Hendry, 1988, Favero and Hendry, 1992, Engle and Hendry, 1993, Psaradakis and Sola, 1996, Jansen and Teräsvirta, 1996, and Hendry and Santos, 2010, including tests based on co-breaking as in Krolzig and Toro, 2002, and Hendry and Massmann, 2007). Here we propose a test of invariance based on step-indicator saturation (denoted SIS), following Castle, Doornik, Hendry, and Pretis (2015). SIS builds on the impulse-indicator saturation (IIS) approach proposed by Hendry, Johansen, and Santos (2008) to detect outliers and shifts in a model.

[2] See https://bankunderground.co.uk/2015/11/20/how-did-the-banks-forecasts-perform-before-during-and-after-the-crisis/.

In IIS, an impulse indicator is created for every observation, thereby saturating the model, but indicators are entered in feasibly large blocks for selection of significant departures. Johansen and Nielsen (2009) derive the properties of IIS for dynamic regression models, which may have unit roots, and show that parameter estimator distributions are almost unaffected despite including T impulse indicators for T observations, and Johansen and Nielsen (2016) derive the distribution of the gauge. Hendry and Mizon (2011a) demonstrate distortions in modelling from not handling outliers, and Castle, Doornik, and Hendry (2016) show that outliers can even be distinguished from non-linear reactions as the latter hold at all observations whereas the former only occur at isolated points. Castle *et al.* (2015) establish that the properties of step-indicator saturation are similar to IIS, but SIS exhibits higher power than IIS when location shifts occur. Ericsson and Reisman (2012) propose combining IIS and SIS in 'super saturation' and Ericsson (2016b) uses IIS to adjust an *ex post* mis-match between the truncated FMI Greenbook 'post-casts' and GDP outcomes.[3]

The invariance test proposed here is the extension to SIS of the IIS-based test of super exogenity in Hendry and Santos (2010). SIS is applied to detect location shifts in models of policy instruments, then any resulting significant step indicators are tested in the model for the target variables of interest. Because the 'size' of a test statistic is only precisely defined for a similar test, and the word is anyway ambiguous in many settings (such as sample size), we use the term 'gauge' to denote the empirical null retention frequency of indicators. The SIS invariance test's gauge is close to its nominal significance level. We also examine its rejection frequencies when parameter invariance does not hold. When the probability of retention of relevant indicators is based on selection, it no longer corresponds to the conventional notion of 'power', so we use the term 'potency' to denote the average non-null retention frequency from selection. However, as IIS can detect failures of invariance from variance shifts in the unmodelled processes, 'super saturation' could help delineate that source of failure. Neither test requires *ex ante* knowledge of the timings, signs, numbers or magnitudes of location shifts, as policy-instrument models are selected automatically using *Autometrics*.

We test for location shifts by adding to the candidate variables a complete set of T step indicators denoted $\left\{1_{\{t\leq j\}}, j=1,\ldots,T\right\}$ when the sample size is T, where $1_{\{t\leq j\}}=1$ for observations up to t, and zero otherwise. Using a modified general-to-specific procedure, Castle *et al.* (2015) establish the gauge and the null distribution of the resulting estimator of regression parameters. A two-block process is investigated analytically, where half the indicators are added and all significant indicators recorded, then that half is dropped, and the other half examined: finally, the two retained sets of indicators are combined. The gauge is approximately αT when the nominal significance level of an individual test is α. Hendry, Johansen, and Santos (2008) and Johansen and Nielsen (2009) show that other splits, such as using k splits of size T/k, or unequal splits, do not affect the gauge of IIS.

The invariance test then involves two stages. Denote the $n=n_1+n_2$ variables in the system by $\mathbf{x}_t'=(\mathbf{y}_t' : \mathbf{z}_t')$, where the $\mathbf{z}_t$ are the conditioning variables. Here, we only consider $n_1=1$. In the first stage, SIS is applied to the marginal system for all n_2 conditioning variables, and the associated significant indicators are recorded. Thus, when the intercept and s lags of all n variables $\mathbf{x}_t$ are always retained (i.e., not subject to selection), SIS is applied at significance level α_1 leading to the selection of $m\geq 0$ step indicators:

$$\mathbf{z}_t=\boldsymbol{\psi}_0+\sum_{i=1}^{s}\boldsymbol{\Psi}_i\mathbf{x}_{t-i}+\sum_{j=1}^{m}\boldsymbol{\eta}_{j,\alpha_1}1_{\{t\leq T_j\}}+\mathbf{v}_{2,t} \tag{19}$$

where (19) is selected to be congruent. The coefficients of the significant step indicators are denoted

[3] All of these indicator saturation methods are implemented in the *Autometrics* algorithm in *PcGive* (see Doornik, 2009 and Doornik and Hendry, 2013), which can handle more variables than observations using block path searches with both expanding and contracting phases (see Hendry and Krolzig, 2005, and Doornik, 2007).

$\boldsymbol{\eta}_{j,\alpha_1}$ to emphasize their dependence on α_1 used in selection. Although SIS could be applied to (19) as a system, we will focus on its application to each equation separately. The gauge of step indicators in such individual marginal models is investigated in Castle *et al.* (2015), who show that simulation-based distributions for SIS using *Autometrics* have a gauge, g, close to the nominal significance level α_1. Subsection 4.1 notes their findings on the potency of SIS at the first stage to determine the occurence of location shifts. Subsection 4.2 provides analytic, and §4.3 Monte Carlo, evidence on the gauge of the invariance test. Subsection 4.4 analyses a failure of invariance from a non-constant marginal process. Subsection 4.5 considers the second stage of the test for invariance. Section 5 investigates the gauges and potencies of the proposed automatic test in Monte Carlo experiments for a bivariate data generation process based on §4.4.

4.1 Potency of SIS at stage 1

Potency could be judged by the selected indicators matching actual location shifts within $0,\ \pm1,\ \pm2$ periods. How often a shift at T_1 (say) is exactly matched by the correct single step indicator $1_{\{t\leq T_1\}}$ is important for detecting policy changes (see e.g., Hendry and Pretis, 2016). However, for stage 2, finding that a shift has occurred within a few periods of its actual occurrence could still allow detection of an invariance failure in a policy model.

Analytic power calculations for a known break point exactly matched by the correct step function in a static regression show high power, and simulations confirm the extension of that result to dynamic models. A lower potency from SIS at stage 1 could be due to retained step indicators not exactly matching the shifts in the marginal model, or failing to detect one or more shifts when there are $N > 1$ location shifts. The former is not very serious, as stage 1 is instrumental, so although there will be some loss of potency at stage 2, attentuating the non-centrality parameter relative to knowing when shifts occurred, a perfect timing match is not essential for the test to reject when invariance does not hold. Failing to detect one or more shifts would be more serious as it both lowers potency relative to knowing those shifts, and removes the chance of testing such shifts at stage 2. Retention of irrelevant step indicators not corresponding to shifts in the marginal process, at a rate determined by the gauge g of the selection procedure, will also lower potency but the effect of this is likely to be small for SIS.

Overall, Castle *et al.* (2015) show that SIS has relatively high potency for detecting location shifts in marginal processes at stage 1, albeit within a few periods either side of their starting/ending, from chance offsetting errors. Thus, we now consider the null rejection frequency at stage 2 of the invariance test for a variety of marginal processes using step indicators selected at stage 1.

4.2 Null rejection frequency of the SIS test for invariance at stage 2

The m significant step indicators $\{1_{\{t\leq j\}}\}$ in the equations of (19) are each retained using the criterion:

$$|\mathsf{t}_{\widehat{\eta}_{i,j,\alpha_1}}| > c_{\alpha_1} \tag{20}$$

when c_{α_1} is the critical value for significance level α_1. Assuming all retained step indicators correspond to distinct shifts (or after eliminating any duplicates), collect them in the m vector $\boldsymbol{\iota}_t$, and add that to the model for y_t (written here with one lag):

$$y_t = \gamma_1 + \boldsymbol{\gamma}_2'\mathbf{z}_t + \boldsymbol{\gamma}_3'\mathbf{x}_{t-1} + \boldsymbol{\delta}_{\alpha_1}'\boldsymbol{\iota}_t + \epsilon_t \tag{21}$$

where $\boldsymbol{\delta}_{\alpha_1}' = (\delta_{1,\alpha_1}\ldots\delta_{m,\alpha_1})$, which should be $\mathbf{0}$ under the null, to be tested as an added-variable set in the conditional equation (21) **without selection**, using an F-test, denoted $\mathsf{F}_{\mathsf{Inv}}$, at significance level α_2 which rejects when $\mathsf{F}_{\mathsf{Inv}(\boldsymbol{\delta}=\mathbf{0})} > c_{\alpha_2}$. Under the null of invariance, this $\mathsf{F}_{\mathsf{Inv}}$-test should have an approximate F-distribution, and thereby allow an appropriately sized test. Under the alternative that $\boldsymbol{\delta} \neq \mathbf{0}$, $\mathsf{F}_{\mathsf{Inv}}$ will have power, as discussed in §4.4.

Although (19) depends on the selection significance level, α_1, the null rejection frequency of $\mathsf{F}_{\mathsf{Inv}}$ should not depend on α_1, although too large a value of α_1 will lead to an F-test with large degrees of freedom, and too small α_1 will lead to few, or even no, step indicators being retained from the marginal models. If no step indicators are retained in (19), so (20) does not occur, then $\mathsf{F}_{\mathsf{Inv}}$ must under-reject when the null is true. Otherwise, the main consideration when choosing α_1 is to allow power against reasonable alternatives to invariance by detecting any actual location shifts in the marginal models.

4.3 Monte Carlo evidence on the null rejection frequency

To check that shifts in marginal processes like (19) do not lead to spurious rejection of invariance when super exogeneity holds, the Monte Carlo experiments need to estimate the gauges, g, of the SIS-based invariance test in three states of nature for (19): (A) when there are no shifts (§4.3.1); (B) for a mean shift (§4.3.2); and (C) facing a variance change (§4.3.3). Values of g close to the nominal significance level α_2 and constant across (A)–(C) are required for a similar test. However, since *Autometrics* selection seeks a congruent model, insignificant irrelevant variables can sometimes be retained, and the gauge will correctly reflect that, so usually $\mathsf{g} \geq \alpha_1$.

The simulation DGP is the non-constant bivariate first-order vector autoregression (VAR(1)):

$$\begin{pmatrix} y_t \\ z_t \end{pmatrix} \mid \mathbf{x}_{t-1} \sim \mathsf{IN}_2 \left[\begin{pmatrix} \gamma_1 + \rho\gamma_{2,t} + \boldsymbol{\kappa}'\mathbf{x}_{t-1} \\ \gamma_{2,t} \end{pmatrix}, \sigma_{22} \begin{pmatrix} \sigma_{22}^{-1}\sigma_{11} + \rho^2\theta_{(t)} & \rho\theta_{(t)} \\ \rho\theta_{(t)} & \theta_{(t)} \end{pmatrix} \right] \tag{22}$$

inducing the valid conditional relation:

$$\mathsf{E}\left[y_t \mid z_t, \mathbf{x}_{t-1}\right] = \gamma_1 + \rho\gamma_{2,t} + \boldsymbol{\kappa}'\mathbf{x}_{t-1} + \frac{\rho\sigma_{22}\theta_{(t)}}{\sigma_{22}\theta_{(t)}}\left(z_t - \gamma_{2,t}\right) = \gamma_1 + \rho z_t + \boldsymbol{\kappa}'\mathbf{x}_{t-1} \tag{23}$$

with:

$$\mathsf{V}\left[y_t | z_t, \mathbf{x}_{t-1}\right] = \sigma_{11} + \rho^2\sigma_{22}\theta_{(t)} - \frac{\rho^2\sigma_{22}^2\theta_{(t)}^2}{\sigma_{22}\theta_{(t)}} = \sigma_{11} \tag{24}$$

In (22), let $\gamma_{2,t} = 1 + \lambda 1_{\{t \leq T_1\}}$ and $\theta_{(t)} = 1 + \theta 1_{\{t \leq T_2\}}$, so the marginal process is:

$$z_t = 1 + \pi 1_{\{t \leq T_1\}} + w_t \quad \text{where} \quad w_t \sim \mathsf{IN}\left[0, \sigma_{22}(1 + \theta 1_{\{t \leq T_2\}})\right] \tag{25}$$

As the analysis in Hendry and Johansen (2015) shows, $\mathbf{x}_{t-1}$ can be retained without selection during SIS, so we do not explicitly include dynamics as the canonical case is testing:

$$\mathsf{H}_0\text{: } \pi = 0 \tag{26}$$

in (25). Despite shifts in the marginal process, (23) shows that invariance holds for the conditional model.

In the Monte Carlo, the constant and invariant parameters of interest are $\gamma_1 = 0$, $\rho = 2$, $\boldsymbol{\kappa} = \mathbf{0}$ and $\sigma_{11} = 1$, with $\sigma_{22} = 5$, and $T_1 = T_2 = 0.8T$. Sample sizes of $T = (50, 100, 200)$ are investigated with $M = 10,000$ replications, for both α_1 (testing for step indicators in the marginal) and α_2 (testing invariance in the conditional equation) equal to (0.025, 0.01, 0.005), though we focus on both at 0.01.

4.3.1 Constant marginal

The simplest setting is a constant marginal process, which is (22) with $\pi = \theta = 0$, so the parameters of the conditional model $y_t | z_t$ are $\boldsymbol{\phi}_1' = (\gamma_1; \rho; \sigma_{11}) = (0; 2; 1)$ and the parameters of the marginal are $\boldsymbol{\phi}_{2,t}' = \left(\gamma_{2,t}; \sigma_{22,t}\right) = (1; 5)$. The conditional representation with m selected indicators from (19) is:

$$y_t = \gamma_1 + \rho z_t + \sum_{j=1}^{m} \delta_{j,\alpha_1} 1_{\{t \leq T_j\}} + \epsilon_t \tag{27}$$

and invariance is tested by the $\mathsf{F}_{\mathsf{Inv}}$-statistic of the null $\boldsymbol{\delta}_{\alpha_1} = \mathbf{0}$ in (27).

$\alpha_1 = 0.01 = \alpha_2$	$T = 50$	$T = 100$	$T = 200$
Stage 1 gauge	0.035	0.033	0.044
% no indicators	0.287	0.098	0.019
Stage 2 gauge	0.006	0.009	0.009

Table 5: SIS simulations under the null of super exogeneity for a constant marginal process. '% no indicators' records the percentage of replications in which no step indicators are retained, so stage 2 is redundant. Stage 2 gauge records the probability of the $\mathsf{F}_{\mathsf{Inv}}$-test falsely rejecting for the included step indicators at $\alpha_2 = 0.01$.

Table 5 records the outcomes at $\alpha = 0.01$ facing a constant marginal process, with $s = 4$ lags of (y, z) included in the implementation of (19). Tests for location shifts in marginal processes should not use too low a probability α_1 of retaining step indicators, or else the $\mathsf{F}_{\mathsf{Inv}}$-statistic will have a zero null rejection frequency. For example, at $T = 50$ and $\alpha_1 = 0.01$, under the null, half the time no step indicators will be retained, so only about $0.5\alpha_2$ will be found overall, as simulation confirms. Simulated gauges and nominal null rejection frequencies for $\mathsf{F}_{\mathsf{Inv}}$ were close so long as $\alpha_1 T \geq 2$.

4.3.2 Location shifts in $\{z_t\}$

The second DGP is given by (22) where $\pi = 2, 10$ with $\gamma_{2,t} = 1 + \pi 1_{\{t \leq T_1\}}$, $\theta = 0$ and $\boldsymbol{\kappa} = \mathbf{0}$. Invariance holds irrespective of the location shift in the marginal, so these experiments check that spurious rejection is not induced thereby. Despite large changes in π, when $T > 100$, Table 6 confirms that gauges are close to nominal significance levels. Importantly, the test does not spuriously reject the null, and now is only slightly undersized at $T = 50$ for small shifts, as again sometimes no step indicators are retained.

		$\pi = 2$			$\pi = 10$	
$\alpha_1 = 0.01$	$T = 50$	$T = 100$	$T = 200$	$T = 50$	$T = 100$	$T = 200$
Stage 1 gauge	0.034	0.027	0.043	0.018	0.018	0.035
Stage 1 potency	0.191	0.186	0.205	0.957	0.962	0.965
% no indicators	0.067	0.022	0.000	0.000	0.000	0.000
Stage 2 gauge:	0.009	0.010	0.011	0.010	0.009	0.010

Table 6: SIS simulations under the null of super exogeneity for a location shift in the marginal process. Stage 1 gauge is for retained step indicators at times with no shifts; and stage 1 potency is for when the exact indicators matching step shifts are retained, with no allowance for mis-timing.

4.3.3 Variance shifts in $\{z_t\}$

The third DGP given by (22) allows the variance-covariance matrix to change while maintaining the conditions for invariance, where $\theta = 2, 10$ with $\pi = 0$, $\gamma_{2,t} = 1$ and $\boldsymbol{\kappa} = \mathbf{0}$. Again, gauges are close to nominal significance levels, and again the test does not spuriously reject the null, remaining slightly undersized at $T = 50$ for small shifts.

$\alpha_1 = 0.01$	$\theta = 2$ $T = 50$	$T = 100$	$T = 200$	$\theta = 10$ $T = 50$	$T = 100$	$T = 200$
Stage 1 gauge	0.042	0.051	0.067	0.060	0.083	0.113
Stage 1 potency	0.030	0.030	0.035	0.041	0.043	0.071
% no indicators	0.381	0.135	0.015	0.342	0.091	0.005
Stage 2 gauge	0.006	0.008	0.009	0.006	0.009	0.010

Table 7: SIS simulations under the null of super exogeneity for a variance shift in the marginal process. Legend as for Table 6.

Overall, the proposed $\mathsf{F}_{\mathsf{Inv}}$ test has appropriate empirical null retention frequencies for both constant and changing marginal processes, so we now turn to its ability to detect failures of invariance.

4.4 Failure of invariance

In this section, we derive the outcome for an invariance failure in the conditional model when the marginal process is non-constant due to a location shift, and obtain the non-centrality and approximate power of the invariance test in the conditional model for a single location shift in the marginal.

From (22) when invariance does not hold:

$$\begin{pmatrix} y_t \\ z_t \end{pmatrix} \mid \mathbf{x}_{t-1} \sim \mathsf{IN}_2 \left[\begin{pmatrix} \gamma_1 + \rho\gamma_{2,t} + \boldsymbol{\kappa}'\mathbf{x}_{t-1} \\ \gamma_{2,t} \end{pmatrix}, \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{pmatrix} \right] \tag{28}$$

so letting $\sigma_{12}/\sigma_{22} = \beta$ leads to the conditional relation:

$$\mathsf{E}\left[y_t \mid z_t, \mathbf{x}_{t-1}\right] = \gamma_1 + (\rho - \beta)\gamma_{2,t} + \beta z_t + \boldsymbol{\kappa}'\mathbf{x}_{t-1} \tag{29}$$

which depends on $\gamma_{2,t}$ when $\rho \neq \beta$ and:

$$z_t = \gamma_{2,t} + \boldsymbol{\nu}_{2,t} \tag{30}$$

When the dynamics and timings and forms of shifts in (30) are not known, we model z_t using (19). *Autometrics* with SIS will be used to select significant regressors as well as significant step indicators from the saturating set $\sum_{j=1}^{T} \eta_{j,\alpha_1} 1_{\{t \leq j\}}$.

Although SIS can handle multiple location shifts, for a tractable analysis, we consider the explicit single alternative (a non-zero intercept in (31) would not alter the analysis):

$$\gamma_{2,t} = \pi 1_{\{t \leq T_1\}} \tag{31}$$

Further, we take γ_1 as constant in (29), so that the location shift is derived from the failure of conditioning when $\rho \neq \beta$:

$$\mathsf{E}\left[y_t \mid z_t, \mathbf{x}_{t-1}\right] = \gamma_1 + (\rho - \beta)\pi 1_{\{t \leq T_1\}} + \beta z_t + \boldsymbol{\kappa}'\mathbf{x}_{t-1} \tag{32}$$

The power of a test of the significance of $1_{\{t \leq T_1\}}$ in a scalar process when the correct shift date is known, so selection is not needed, is derived in Castle *et al.* (2015), who show that the test power rises with the magnitude of the location shift and the length it persists, up to half the sample. Their results apply to testing $\pi = 0$ in (31) and would also apply to testing $(\rho - \beta)\pi = 0$ in (32) when $1_{\{t \leq T_1\}}$ is known or is correctly selected at Stage 1. Subsection 4.5 considers the impact on the power of the invariance test of $(\rho - \beta)\pi = 0$ of needing to discover the shift indicator at Stage 1. Castle *et al.* (2015) also examine the effects on the potency at Stage 1 of selecting a step indicator that does not precisely match the location shift in the DGP, which could alter the power at Stage 2.

4.5 Second-stage test

The gauge of the $\mathsf{F}_{\mathsf{Inv}}$ test at the second stage conditional on locating the relevant step indicators corresponding exactly to the shifts in the marginal was calculated above. A relatively loose α_1 will lead to retaining some 'spurious' step indicators in the marginal, probably lowering potency slightly. In practice, there could be multiple breaks in different marginal processes at different times, which may affect one or more $z_{j,t}$, but little additional insight is gleaned compared to the one-off break in (31), since the proposed test is an F-test on all retained step indicators, so does not assume any specific shifts at either stage. The advantage of using the explicit alternative in (31) is that approximate analytic calculations are feasible. We only consider a bivariate VAR explicitly, where $(\rho - \beta) = 0$ in (29) under the null that the conditional model of $\{y_t\}$ is invariant to the shift in z_t.

Let SIS selection applied to the marginal model for z_t yield a set of step indicators $\mathcal{S}_{\alpha_1}$ defined by:

$$\mathcal{S}_{\alpha_1} = \left\{ \mathsf{t}^2_{\eta_{i,\alpha_1}=0} > c^2_{\alpha_1} \right\} \tag{33}$$

which together entail retaining $\left\{1_{\{t\leq i\}}, i = 1, \ldots, m\right\}$. Combine these significant step indicators in a vector $\boldsymbol{\iota}_t$, and add $\boldsymbol{\iota}_t$ to the assumed constant relation:

$$y_t = \gamma_1 + \beta z_t + \boldsymbol{\kappa}'\mathbf{x}_{t-1} + \nu_{1,t} \tag{34}$$

to obtain the test regression written as:

$$y_t = \mu_0 + \mu_1 z_t + \boldsymbol{\mu}_3'\mathbf{x}_{t-1} + \boldsymbol{\delta}'\boldsymbol{\iota}_t + e_t \tag{35}$$

As with IIS, a difficulty in formalizing the analysis is that the contents of $\boldsymbol{\iota}_t$ vary between draws, as it matters how close the particular relevant step indicators retained are to the location shift, although which specific irrelevant indicators are retained will not matter greatly. However, Hendry and Pretis (2016) show that the main reason SIS chooses indicators that are incorrectly timed is that the shift is either initially 'camouflaged' by opposite direction innovation errors, or same-sign large errors induce an earlier selection. In both cases, such mistakes should only have a small impact on the second-stage test power as simulations confirm. Failing to detect a shift in the marginal model will lower power when that shift in fact leads to a failure of invariance in the conditional model.

4.6 Static bivariate case

We first consider the impact of mis-timing the selection of an indicator for a location shift in the conditional process in (29) derived from the non-invariant system (28) when $\boldsymbol{\kappa} = \mathbf{0}$. The conditional relation is written for $(\rho - \beta)\pi = \phi$ as:

$$y_t = \gamma_1 + \beta z_t + \phi 1_{\{t\leq T_1\}} + \epsilon_t \tag{36}$$

where $\epsilon_t \sim \mathsf{IN}\left[0, \sigma^2_\epsilon\right]$, which lacks invariance to changes in π in the marginal model (19):

$$z_t = \pi 1_{\{t\leq T_1\}} + \nu_{2,t} \tag{37}$$

However, (36) is modelled by:

$$y_t = \mu_0 + \mu_1 z_t + \delta 1_{\{t\leq T_0\}} + e_t \tag{38}$$

when $T_0 \neq T_1$.

As SIS seeks the best matching step indicator for the location shift, any discrepancy between $1_{\{t\leq T_1\}}$ and $1_{\{t\leq T_0\}}$ is probably because the values of $\{\epsilon_t\}$ between T_0 and T_1 induced the mistaken choice. Setting $T_1 = T_0 + T_+$ where $T_+ > 0$ for a specific formulation, then:

$$e_t = (\gamma_1 - \mu_0) + ((\rho - \mu_1)\pi - \delta) 1_{\{t\leq T_0\}} + (\rho - \mu_1)\pi 1_{\{T_0+1\leq t\leq T_1\}} + \epsilon_t + (\beta - \mu_1)\nu_{2,t} \tag{39}$$

For observations beyond T_1:

$$e_t = (\gamma_1 - \mu_0) + (\beta - \mu_1)\,\nu_{2,t} + \epsilon_t$$

so that:

$$\mathsf{E}\left[\sum_{t=T_1+1}^{T} e_t^2\right] = (T - T_1)\left[(\gamma_1 - \mu_0)^2 + (\beta - \mu_1)^2\,\sigma_{v_2}^2 + \sigma_\epsilon^2\right] \tag{40}$$

which would be large for $\gamma_1 \neq \mu_0$ and $\beta \neq \mu_1$. To a first approximation, least-squares selection would drive estimates such that the first two terms in (40) vanish. If so, for $t \leq T_0$, (39) becomes:

$$e_t = (\phi - \delta)\,1_{\{t\leq T_0\}} + \epsilon_t \tag{41}$$

and between T_0 and T_1:

$$e_t = \phi 1_{\{T_0+1\leq t\leq T_1\}} + \epsilon_t \tag{42}$$

suggesting SIS would find $\delta \approx \phi$, and that chance draws of $\{\epsilon_t\}$ must essentially offset $\phi 1_{\{T_0+1\leq t\leq T_1\}}$ during the non-overlapping period $T_0 + 1$ to T_1. In such a setting:

$$\mathsf{E}\left[\frac{1}{T}\sum_{t=1}^{T} e_t^2\right] \approx \sigma_\epsilon^2 + \frac{(T_1 - T_0)}{T}\phi^2$$

so that the estimated equation error variance will not be greatly inflated by the mis-match in timing, and the rejection frequency of a t-test of H_0: $\delta = 0$ will be close to that of the corresponding test of H_0: $\phi = 0$ in (36) despite the selection of the relevant indicator by SIS.

The power of the $\mathsf{F}_{\mathsf{Inv}}$-test of H_0: $\delta = 0$ in (38) depends on the strength of the invariance violation, $\rho - \beta$; the magnitude of the location shift, π, both directly and through its detectability in the marginal model, which in turn depends on α_1; the sample size T; the number of periods T_1 affected by the location shift; the number of irrelevant step indicators retained (which will affect the test's degrees of freedom, again depending on α_1); how close the selected step shift is to matching the DGP shift; and on α_2. The properties are checked by simulation below, and can be contrasted in experiments with the optimal, but generally infeasible, test based on adding the precisely correct indicator $1_{\{t\leq T_1\}}$, discussed above. Section 5 reports the simulation outcomes.

5 Simulating the powers of the SIS invariance test

The simulation analyses used the bivariate relationship in §4.3 for violations of super exogeneity due to a failure of weak exogeneity under non-constancy in:

$$\begin{pmatrix} y_t \\ z_t \end{pmatrix} \sim \mathsf{IN}_2\left[\begin{pmatrix} \gamma_1 + \rho\gamma_{2,t} \\ \gamma_{2,t} \end{pmatrix}, \begin{pmatrix} 21 & 10 \\ 10 & 5 \end{pmatrix}\right] \tag{43}$$

where $\gamma_1 = 2$ and $\omega^2 = \sigma_{11} - \sigma_{12}^2/\sigma_{22} = 1$, but $\beta = \sigma_{12}/\sigma_{22} = 2 \neq \rho$, with a level shift at T_1 in the marginal:

$$\gamma_{2,t} = \pi 1_{\{t>T_1\}} \quad \text{so} \quad \gamma_{1,t} = \gamma_1 + \rho\pi 1_{\{t>T_1\}} \tag{44}$$

We vary $d = \pi/\sqrt{\sigma_{22}}$ over the values 1, 2, 2.5, 3 and 4; ρ over 0.75, 1, 1.5 and 1.75 when $\beta = 2$, reducing the extent of departure from weak exogeneity; for a sample size of $T = 100$ with a break point at $T_1 = 80$; and significance levels $\alpha_1 = 0.01$ and $\alpha_2 = 0.01$ in the marginal and conditional, with $M = 1,000$ replications.

5.1 Optimal infeasible indicator-based F-test

The optimal infeasible F-test with a known break location in the marginal process is computable in simulations. Table 8 reports the rejections of invariance, which are always high for large shifts, and fall as departures from weak exogeneity decrease. The empirical rejection frequencies approximate maximum achievable power for this type of test. When the location shift is larger than $2.5\sqrt{\sigma_{22}}$, the correct step indicator is almost always significant in the conditional model even for relatively small vaues of $(\rho - \beta)$.

$d : \rho$	0.75	1	1.5	1.75
1	1.000	1.000	0.886	0.270
2	1.000	1.000	1.000	0.768
2.5	1.000	1.000	1.000	0.855
3	1.000	1.000	1.000	0.879
4	1.000	1.000	1.000	0.931

Table 8: Power of the optimal infeasible F-test for a failure of invariance using a known step indicator for $\alpha_2 = 0.01$ at $T_1 = 80$, $T = 100$, $M = 1,000$.

5.2 Power of the SIS-based test

Table 9 records the Stage 1 gauge and potency at different levels of location shift (d) and departures from super exogeneity $(\rho - \beta)$. The procedure is slightly over-gauged at Stage 1 for small shifts, when its potency is also low, and both gauge and potency are unaffected by the magnitude of $(\rho - \beta)$, whereas Stage 1 potency rises rapidly with d.

	Stage 1 gauge				Stage 1 potency			
$d : \rho$	0.75	1	1.5	1.75	0.75	1	1.5	1.75
1	0.038	0.040	0.041	0.039	0.231	0.223	0.227	0.204
2	0.029	0.028	0.030	0.030	0.587	0.575	0.603	0.590
2.5	0.026	0.026	0.025	0.025	0.713	0.737	0.730	0.708
3	0.023	0.023	0.024	0.025	0.820	0.813	0.803	0.817
4	0.020	0.021	0.020	0.022	0.930	0.930	0.922	0.929

Table 9: Stage 1 gauge and potency at $\alpha_1 = 0.01$ for $T_1 = 80$, $T = 100$, $M = 1,000$ and $\beta = 2$.

Table 10 records Stage 2 power for the three values of α_1. It shows that for a failure of invariance, even when $\rho - \beta = 0.25$, test power can increase at tighter Stage 1 significance levels, probably by reducing the retention rate of irrelevant step indicators. Comparing the central panel with the matching experiments in Table 8, there is remarkably little loss of power from selecting the indicator by SIS at Stage 1, rather than knowing it, except at the smallest values of d.

6 Application to the small artificial-data policy model

To simulate a case of invariance failure from a policy change, which could be checked by $\mathsf{F}_{\mathsf{Inv}(\boldsymbol{\delta}=\mathbf{0})}$ in-sample, followed by forecast failure, we splice the two scenarios from Section 3.3 sequentially in the order of the 100 observations in panels (III+IV) then those used for panels (I+II), creating a sample of $T = 200$. Next, we estimate (5) with SIS, retaining the policy variable, and test the significance of the selected step-indicators in (6). At Stage 1, using $\alpha_1 = 0.001$, as the model is mis-specified and the

	$\alpha_1 = 0.025$				$\alpha_1 = 0.01$				$\alpha_1 = 0.005$			
$d:\rho$	0.75	1	1.5	1.75	0.75	1	1.5	1.75	0.75	1	1.5	1.75
1	0.994	0.970	0.378	0.051	0.918	0.908	0.567	0.127	0.806	0.798	0.535	0.123
2	1.000	1.000	0.966	0.355	1.000	1.000	0.994	0.604	0.999	0.999	0.997	0.653
2.5	1.000	1.000	0.995	0.499	1.000	1.000	0.999	0.744	1.000	0.999	0.998	0.786
3	1.000	1.000	0.999	0.594	1.000	1.000	1.000	0.821	0.999	0.999	1.000	0.861
4	1.000	1.000	0.998	0.712	1.000	1.000	0.999	0.912	0.999	1.000	0.999	0.942

Table 10: Stage 2 power for a failure of invariance at $\alpha_2 = 0.01$, $T_1 = 80$, $T = 100$, and $M = 1,000$.

sample is $T = 189$ keeping the last 11 observations for the forecast period, two indicators are selected. Testing these in (6) yields $\mathsf{F}_{\mathsf{Inv}(\boldsymbol{\delta}=\mathbf{0})}(2, 186) = 13.68^{**}$, strongly rejecting invariance of the parameters of the model for y_t to shifts in the model of x_t.

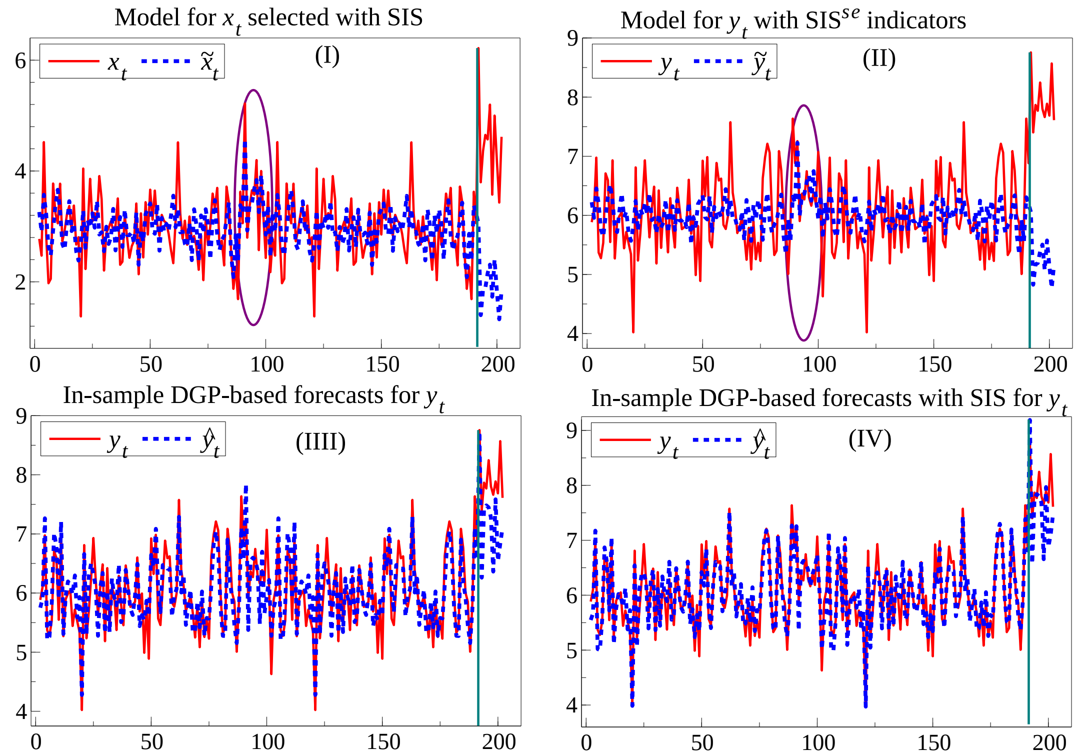

Figure 2: (I) Forecast failure for x_t by $\widetilde{x}_t$ even with SIS; (II) Forecast failure in $\widetilde{y}_t$ even augmented by the SIS indicators selected from the margin model for x_t; (III) Smaller forecast failure from $\widehat{y}_t$ based on the in-sample DGP; (IV) Least forecast failure from $\widehat{y}_t$ based on the in-sample DGP with SIS.

Figure 2 reports the outcome graphically, where the ellipses show the period with the ealier break in the DGP without a location shift but with the policy change. Panel (*I*) shows the time series for x_t with the fitted and forecast values, denoted $\widetilde{x}_t$, from estimating the agency's model with SIS which delivered the indicators for testing invariance. Panel (*II*) shows the outcome for y_t after adding the selected indicators denoted SIS^{se} from the marginal model for x_t, which had earlier rejected invariance. Panel (*III*) reports the outcome for a setting where the invariance failure led to an improved model, which here coincides with the in-sample DGP. This greatly reduces the later forecast errors and forediction failures. Finally, Panel (*IV*) augments the estimated in-sample DGP equation (with all its regressors retained) by selecting using SIS at $\alpha = 0.01$. This further reduces forecast failure, although constancy can still be rejected from the unanticipated location shift. If the invariance rejection had led to the development of an improved model, better forecasts, and hopefully improved foredictions and policy decisions, would have resulted. When the policy model is not known publicly (as with MPC decisions), the agency alone can conduct these tests.

6.1 Multiplicative indicator saturation

In the preceding example, the invariance failure was detectable by SIS because the policy change created a location shift by increasing z_t by δ. A zero-mean shift in a policy-relevant derivative, would not be detected by SIS, but could be by multiplicative indicator saturation (MIS) proposed in Ericsson (2012). MIS interacts step indicators with variables as in $d_{j,t} = z_t 1_{\{j \leq t\}}$, so $d_{j,t} = z_t$ when $j \leq t$ and is zero otherwise. Kitov and Tabor (2015) have investigated its performance in detecting changes in parameters in zero-mean settings by extensive simulations. Despite the very high dimensionality of the resulting parameter space, they find MIS has a gauge close to the nominal significance level for suitably tight α, and has potency to detect such parameter changes. As policy failure will occur after a policy-relevant parameter shifts, advance warning thereof would be invaluable. Even though the above illustration detected a failure of invariance, it did not necessarily entail that policy-relevant parameters had changed. We now apply MIS to the $T = 200$ artificial data example in section 6 to ascertain whether such a change could be detected, focusing on potential shifts in the coefficient of z_{t-1} in (6).

Selecting at $\alpha = 0.005$ as there are more then 200 candidate variables yielded:

$$y_t = -\underset{(0.27)}{1.12}\, z_{t-1}1_{\{t\leq 90\}} + \underset{(0.27)}{1.17}\, z_{t-1}1_{\{t\leq 100\}} + \underset{(0.045)}{6.01} - \underset{(0.2)}{0.857}\, z_{t-1} \tag{45}$$

with $\widehat{\sigma} = 0.60$. Thus, the in-sample shift of -1 in $\lambda_1\theta_1$ is found over $t = 90, \ldots, 100$, warning of a lack of invariance in the key policy parameter from the earlier policy change, although that break is barely visible in the data, as shown by the ellipse in Figure 2 (*II*). To understand how MIS is able to detect the parameter change, consider knowing where the shift ocurred and splitting your data at that point. Then you would be startled if fitting your correctly specified model separately to the different subsamples did not deliver the appropriate estimates of their DGP parameters. Choosing the split by MIS will add variability, but the correct indicator, or one close to it, should accomplish the same task.

7 How to improve future forecasts and foredictions

Scenarios above treated direct forecasts and those derived from foredictions as being essentially the same. When a major forecast error occurs, the agency can use a robust forecasting device such as an intercept correction (IC), or differencing the forecasts, to set them 'back on track' for the next period. However, the foredictions that led to the wrong policy implementation cannot be fixed so easily, even if the agency's next narrative alters the story. In our example, further increases in δ will induce greater forecast failure if the policy model is unchanged: viable *policy* requires invariance of the model to the policy change. Nevertheless, there is a partial 'fix' to the forecast failure and policy invalidity. If the lack of invariance is invariant, so policy shifts change the model's parameters in the same way each time as in (14), the shift associated with a past policy change can be added as an IC to a forecast based on a later policy shift. We denote such an IC by SIS^{se}_{IC}, which has the advantage that it can be implemented before experiencing forecast failure. This is shown in Figure 3 (I) & (II), focusing on the last 50 periods, where the policy change coincides with the location shift at observation 191. The first panel records the forecasts for a model of y_t which includes the SIS^{se} indicator for the period of the first forecast failure, and also includes SIS^{se}_{IC} as an imposed IC, from $T = 191$. Had a location shift not also occurred, SIS^{se}_{IC} would have corrected the forecast for the lack of invariance, and could have been included in the policy analysis and the associated foredictions.

Figure 3 also shows how effective a conventional IC is in the present context after the shift has occurred. The IC is a step indicator with a value of unity from observation $t = 191$ onwards when the forecast origin is $T = 192$, so one observation on the forecast error is used to estimate the location shift. Compared to the massive forecast failure seen for the models of y_t in Figure 2 (I) & (II), neither of the sets of forecast errors in Figure 3 (III) & (IV) fails a constancy test ($\mathsf{F}_{\mathsf{Chow}}(10, 188) = 0.34$ and $\mathsf{F}_{\mathsf{Chow}}(10, 185) = 0.48$). The IC alone corrects most of the forecast failure, but as (IV) shows, SIS improves the in-sample tracking by correcting the earlier location-shift induced failure and improves the resulting forecasts.[4]

In real-time forecasting, these two steps could be combined, using SIS^{se}_{IC} as the policy is implemented, followed by an IC one-period later when the location shift materialises, although a further policy change is more than likely in that event. Here, the mis-specified econometric models of the relationships between the variables are unchanged, but the forecasts are very different: successful forecasts do not imply correct models.

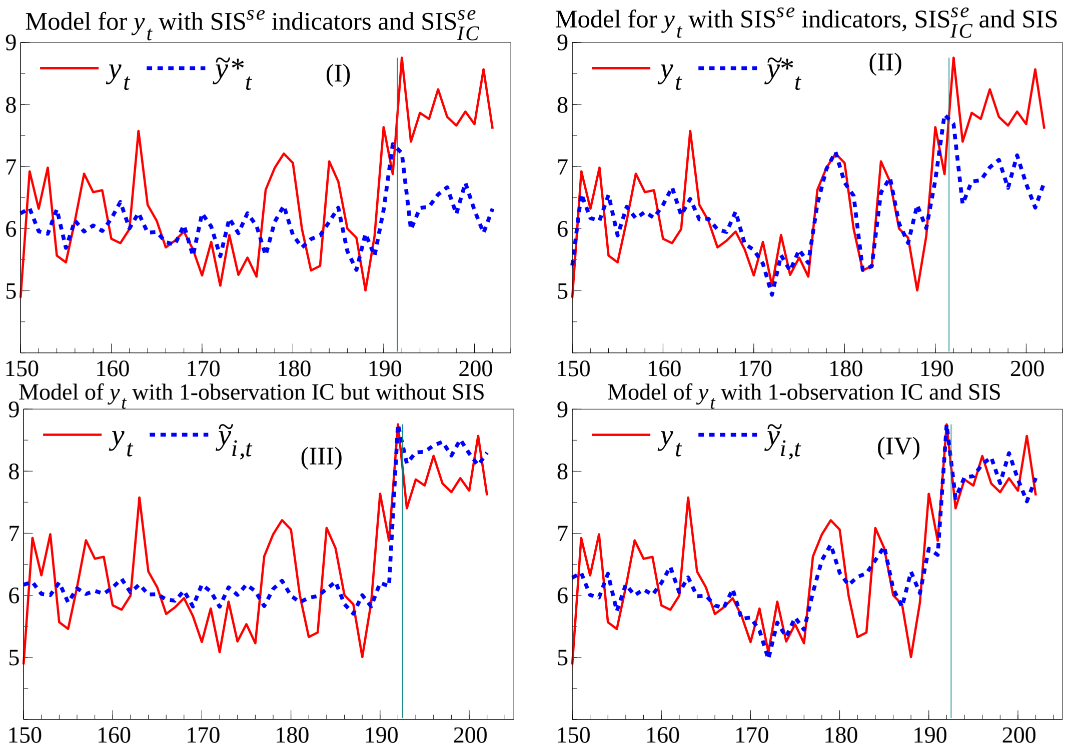

Figure 3: (I) Forecasts for y_t by $\widetilde{y}^*_t$, just using SIS^{se} for the first policy-induced shift and SIS^{se}_{IC} at observation 191; (II) Forecasts for y_t by $\widetilde{y}^*_t$ also with SIS in-sample; (III) Forecasts from $T = 192$ for y_t by $\widetilde{y}_{i,t}$ with a 1-observation IC but without SIS; (IV) Forecasts for y_t by $\widetilde{y}_{i,t}$ also with SIS.

[4] As noted above, the lagged impact of the policy change causes x_{191} to overshoot, so $\widetilde{x}_{i,t}$ is somewhat *above* x_t over the forecast horizon, albeit a dramatic improvement over Figure 2 (I): using 2-periods to estimate the IC solves that.

8 Conclusion

We considered two potential implications of forecast failure in a policy context, namely forediction failure and policy invalidity. While an empirical forecasting model cannot necessarily be rejected following forecast failure, when the forecasts derived from the narratives of a policy agency are very close to the model's forecasts, as Ericsson (2016b) showed was true for the FOMC minutes, then forecast failure entails forediction failure. Consequently, the associated narrative and any policy decisions based thereon also both fail. A taxonomy of the sources of forecast errors showed what could be inferred from forecast failure, and was illustrated by a small artificial-data policy model.

A test for invariance and the validity of policy analysis was proposed by selecting shifts in all marginal processes using step-indicator saturation and checking their significance in the conditional model. The test was able to detect failures of invariance when weak exogeneity failed and the marginal processes changed from a location shift. Compared to the nearest corresponding experiment in Hendry and Santos (2010), the power of $\mathsf{F}_{\mathsf{Inv}}$ is considerably higher for SIS at $\alpha_2 = 0.01$ than IIS at $\alpha_2 = 0.025$ (both at $\alpha_1 = 0.025$) as shown in Figure 4.

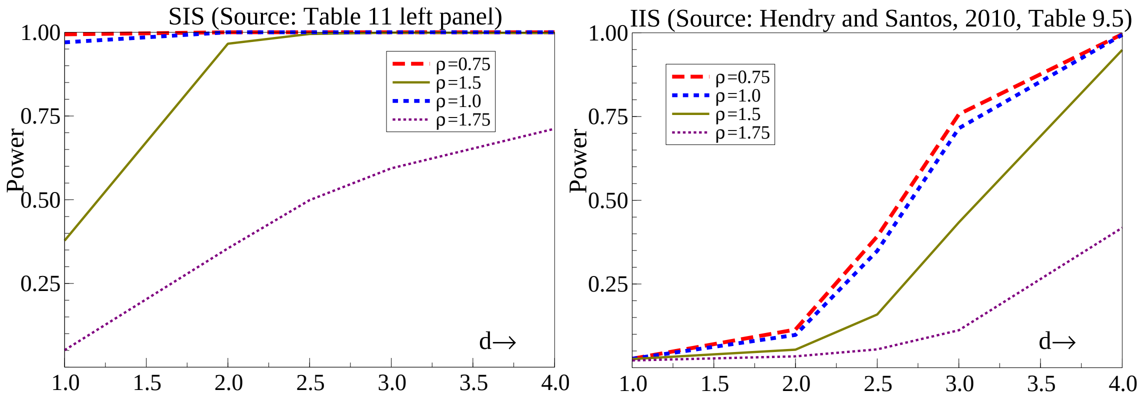

Figure 4: Comparison of the power of SIS with IIS.

A test rejection outcome by $\mathsf{F}_{\mathsf{Inv}}$ indicates a dependence between the conditional model parameters and those of the marginals, warning about potential mistakes from using the conditional model to predict the outcomes of policy changes that alter the marginal processes by location shifts, which is a common policy scenario. Combining these two features of forecast failure with non-invariance allows forediction failure and policy invalidity to be established when they occur. Conversely, learning that the policy model is not invariant to policy changes could lead to improved models, and we also showed a 'fix' that could help mitigate forecast and policy failure.

While all the derivations and Monte Carlo experiments here have been for 1-step forecasts from static regression equations, a single location shift and a single policy change, the general nature of the test makes it applicable when there are multiple breaks in several marginal processes, perhaps at different times. Generalizations to dynamic equations, to conditional systems, and to other non-stationary settings, probably leading to more approximate null rejection frequencies, are the focus of our present research.

References

Akram, Q. F. and R. Nymoen (2009). Model selection for monetary policy analysis: How important is empirical validity? *Oxford Bulletin of Economics and Statistics 71*, 35–68.

Bank of England (2015). *Inflation Report, August, 2015*. London: Bank of England Monetary Policy Committee.

Cartwright, N. (1989). *Nature's Capacities and their Measurement*. Oxford: Clarendon Press.

Castle, J. L., J. A. Doornik, and D. F. Hendry (2016). Robustness and model selection. Unpublished paper, Economics Department, University of Oxford.

Castle, J. L., J. A. Doornik, D. F. Hendry, and F. Pretis (2015). Detecting location shifts during model selection by step-indicator saturation. *Econometrics 3(2)*, 240–264.

Castle, J. L. and D. F. Hendry (2011). On not evaluating economic models by forecast outcomes. *Istanbul University Journal of the School of Business Administration 40*, 1–21.

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

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

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

Clements, M. P. and J. J. Reade (2016). Forecasting and forecast narratives: The Bank of England inflation reports. Discussion paper, Economics Department, University of Reading.

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

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

Doornik, J. A. and D. F. Hendry (2013). *Empirical Econometric Modelling using PcGive: Volume I.* (7th ed.). London: Timberlake Consultants Press.

Ellison, M. and T. J. Sargent (2012). A defense of the fomc. *International Economic Review 53*(4), 1047–1065.

Engle, R. F. and D. F. Hendry (1993). Testing super exogeneity and invariance in regression models. *Journal of Econometrics 56*, 119–139.

Engle, R. F., D. F. Hendry, and J.-F. Richard (1983). Exogeneity. *Econometrica 51*, 277–304.

Ericsson, N. R. (2012). Detecting crises, jumps, and changes in regime. Working paper, Federal Reserve Board of Governors, Washington, D.C.

Ericsson, N. R. (2016a). Economic forecasting in theory and practice: An interview with David F. Hendry. *International Journal of Forecasting*, Forthcoming.

Ericsson, N. R. (2016b). Eliciting GDP forecasts from the FOMC's minutes around the financial crisis. *International Journal of Forecasting 32*, 571–583.

Ericsson, N. R. and E. L. Reisman (2012). Evaluating a global vector autoregression for forecasting. *International Advances in Economic Research 18*, 247–258.

Favero, C. and D. F. Hendry (1992). Testing the Lucas critique: A review. *Econometric Reviews 11*, 265–306.

Genberg, H. and A. B. Martinez (2014). On the accuracy and efficiency of IMF forecasts: A survey and some extensions. IEO Background Paper BP/14/04, Independent Evaluation Office of the International Monetary Fund.

Hendry, D. F. (1988). The encompassing implications of feedback versus feedforward mechanisms in econometrics. *Oxford Economic Papers 40*, 132–149.

Hendry, D. F. (2001). How economists forecast. In D. F. Hendry and N. R. Ericsson (Eds.), *Understanding Economic Forecasts*, pp. 15–41. Cambridge, Mass.: MIT Press.

Hendry, D. F. (2004). Causality and exogeneity in non-stationary economic time series. In A. Welfe (Ed.), *New Directions in Macromodelling*, pp. 21–48. Amsterdam: North Holland.

Hendry, D. F. (2006). Robustifying forecasts from equilibrium-correction models. *Journal of Econometrics 135*, 399–426.

Hendry, D. F. and S. Johansen (2015). Model discovery and Trygve Haavelmo's legacy. *Econometric Theory 31*, 93–114.

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

Hendry, D. F. and H.-M. Krolzig (2005). The properties of automatic Gets modelling. *Economic Journal 115*, C32–C61.

Hendry, D. F. and M. Massmann (2007). Co-breaking: Recent advances and a synopsis of the literature. *Journal of Business and Economic Statistics 25*, 33–51.

Hendry, D. F. and G. E. Mizon (2011a). Econometric modelling of time series with outlying observations. *Journal of Time Series Econometrics 3 (1)*, DOI: 10.2202/1941–1928.1100.

Hendry, D. F. and G. E. Mizon (2011b). What needs re-thinking in macroeconomics? *Global Policy 2*, 176–183.

Hendry, D. F. and G. E. Mizon (2012). Open-model forecast-error taxonomies. In X. Chen and N. R. Swanson (Eds.), *Recent Advances and Future Directions in Causality, Prediction, and Specification Analysis*, pp. 219–240. New York: Springer.

Hendry, D. F. and G. E. Mizon (2014). Unpredictability in economic analysis, econometric modeling and forecasting. *Journal of Econometrics 182*, 186–195.

Hendry, D. F. and F. Pretis (2016). Quantifying the uncertainty around break dates in models using indicator saturation. Working paper, Economics Department, Oxford University.

Hendry, D. F. and C. Santos (2010). An automatic test of super exogeneity. In M. W. Watson, T. Bollerslev, and J. Russell (Eds.), *Volatility and Time Series Econometrics*, pp. 164–193. Oxford: Oxford University Press.

Independent Evaluation Office, I. (2014). IMF forecasts: Process, quality, and country perspectives. Technical report, International Monetary Fund, Washington.

Jansen, E. S. and T. Teräsvirta (1996). Testing parameter constancy and super exogeneity in econometric equations. *Oxford Bulletin of Economics and Statistics 58*, 735–763.

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

Johansen, S. and B. Nielsen (2016). Asymptotic theory of outlier detection algorithms for linear time series regression models. *Scandinavian Journal of Statistics 43*, 321–348.

Kitov, O. I. and M. N. Tabor (2015). Detecting structural changes in linear models: A variable selection approach using multiplicative indicator saturation. Unpublished paper, University of Oxford.

Krolzig, H.-H. and J. Toro (2002). Testing for super-exogeneity in the presence of common deterministic shifts. *Annales d'Economie et de Statistique 67/68*, 41–71.

Leamer, E. E. (2009). *Macroeconomic Patterns and Stories*. SBerlin: Springer-Verlag.

Pagan, A. R. (2003). Report on modelling and forecasting at the Bank of England. *Bank of England Quarterly Bulletin*, Spring. http://www.bankofengland.co.uk/archive/Documents/historicpubs/qb/2003/qb030106.pdf.

Psaradakis, Z. and M. Sola (1996). On the power of tests for superexogeneity and structural invariance. *Journal of Econometrics 72*, 151–175.

Romer, C. D. and D. H. Romer (2008). The fomc versus the staff: where can monetary policymakers add value? *American Economic Review 98*(2), 230–35.

Sinclair, T., P.-L. Tien, and E. Gamber (2016). Do fed forecast errors matter? Technical report, CAMA Working Paper.

Siviero, S. and D. Terlizzese (2001). Macroeconomic forecasting: debunking a few old wives' tales. Discussion paper 395, Research Department, Banca d'Italia.

Spanos, A. (2007). Curve-fitting, the reliability of inductive inference and the error-statistical approach. *Philosophy of Science 74*, 1046–1066.

Stekler, H. O. and H. Symington (2016). Evaluating qualitative forecasts: The FOMC minutes, 2006–2010. *International Journal of Forecasting 32*, 559–570.

Stenner, A. J. (1964). On predicting our future. *Journal of Philosophy 16*, 415–428.

Zhang, K., J. Zhang, and B. Schölkopf (2015). Distinguishing cause from effect based on exogeneity. *http://arxiv.org/abs/1504.05651*.