International Journal of Forecasting 34 (2018) 119–135

Contents lists available at ScienceDirect

International Journal of Forecasting

journal homepage: www.elsevier.com/locate/ijforecast

Deciding between alternative approaches in macroeconomics

David F. Hendry

Economics Department and Institute for New Economic Thinking at the Oxford Martin School, University of Oxford, UK

ARTICLE INFO

Keywords:
Model selection
Theory retention
Location shifts
Indicator saturation
Autometrics

ABSTRACT

Macroeconomic time-series data are aggregated, inaccurate, non-stationary, collinear and rarely match theoretical concepts. Macroeconomic theories are incomplete, incorrect and changeable: location shifts invalidate the law of iterated expectations and 'rational expectations' are then systematically biased. Empirical macro-econometric models are non-constant and mis-specified in numerous ways, so economic policy often has unexpected effects, and macroeconomic forecasts go awry. In place of using just one of the four main methods of deciding between alternative models, theory, empirical evidence, policy relevance and forecasting, we propose nesting 'theory-driven' and 'data-driven' approaches, where theory-models' parameter estimates are unaffected by selection despite searching over rival candidate variables, longer lags, functional forms, and breaks. Thus, theory is retained, but not imposed, so can be simultaneously evaluated against a wide range of alternatives, and a better model discovered when the theory is incomplete.

1. Introduction

All macroeconomic theories are incomplete, incorrect and changeable; all macroeconomic time-series data are aggregated, inaccurate, non-stationary and rarely match theoretical concepts; all empirical macro-econometric models are non-constant, and mis-specified in numerous ways; macroeconomic forecasts often go awry: and economic policy often has unexpected effects different from prior analyses; so how should we decide between alternative approaches to modelling macroeconomies?

Historically, the main justification of empirical macro-econometric evidence has been conformity with conventionally-accepted macroeconomic theory: internal credibility as against verisimilitude. Yet in most sciences, theory consistency and verisimilitude both matter and neither can claim precedence: why is economics different? Part of the justification for a 'theory-driven' stance in economics is the manifest inadequacy of short, interdependent, non-stationary and heterogeneous time-series data, often subject to extensive revision. If data are unreliable, it could be argued that perhaps it is better to trust the theory. But macroeconomic theories are inevitably abstract, and usually ignore non-stationarity and aggregation over heterogeneous entities, so are bound to be incorrect and incomplete. Moreover, theories have evolved greatly, and many previous economic analyses have been abandoned, so it is self-contradictory to justify an empirical model purely by an invalid theory that will soon be altered. It is unclear why an incorrect and mutable theory is more reliable than data that can become more accurate over time.

The prevalence of the constellation of non-stationarity, endogeneity, potential lack of identification, inaccurate data and collinearity, have culminated in a belief that 'data-driven' is tantamount to 'data mining' and can produce almost any desired result—but unfortunately so can theory choice by claiming to match idiosyncratic choices of 'stylized facts', that are usually neither facts nor constant. The preference of theory over evidence may also be due to a mistaken conflation of economic-theory models of human behaviour with the data generation process (DGP): there

E-mail address: david.hendry@nuffield.ox.ac.uk.

https://doi.org/10.1016/j.ijforecast.2017.09.003

is a huge gap between abstract theory and non-stationary evidence, inadequately finessed by simply asserting that the model is the mechanism—while *ceteris paribus* is easily assumed in theories, it is rarely achieved empirically (see Boumans, 1999). Indeed, Hendry and Mizon (2014) highlight fundamental failures in the mathematical basis of inter-temporal macroeconomic theory in wide-sense non-stationary economies, namely where the distributions of all economic variables are not the same at all points in time. Such shifts help explain the 'breakdown' of the Bank of England quarterly econometric model (BEQEM),[1] the empirical rejections of 'rational expectations' models in Castle, Doornik, Hendry, and Nymoen (2014), and the failures of economic forecasting discussed in Clements and Hendry (1998).

Finally, there is a false belief that data-based model selection is a subterfuge of scoundrels—rather than the key to understanding the complexities of macroeconomies. All decisions about a theory formulation, its evidential database, its empirical implementation and its evaluation involve selection, though such decisions are often either ignored or camouflaged: **selection is unavoidable and ubiquitous**. Building on that insight, a replacement approach is proposed in Hendry and Doornik (2014) and Hendry and Johansen (2015) that retains theories while selecting across a large range of alternatives, including any number of shifts of unknown magnitudes and signs at unknown times. Thus, instead of simply adopting one of the four conventional methods of deciding between alternative models, namely macroeconomic theory, empirical evidence, policy relevance and forecasting, all of which transpire to be inadequate individually in the face of the complexities of macroeconomies observed through their resulting aggregate time series, we propose an extension of encompassing (see Mizon and Richard, 1986) that nests 'theory-driven' and 'data-driven' approaches, as in Hendry and Johansen (2015). Theory insights can be retained **unaffected** by data-based model selection even when searching over many rival candidate variables, longer lags, non-linear functional forms, and structural breaks. Despite commencing from very general specifications, possibly with more variables, N, than observations T, multi-path search algorithms such as *Autometrics* (see Doornik, 2009) can control retention rates of irrelevant variables at low levels using stringent critical values, yet automatically retain all theory-based variables, irrespective of their significance. When the embedded theory is correct and complete, the distributions of its estimated parameters are *identical* to those obtained by directly fitting it to data. Conversely, because the theory model is retained but not imposed, when it is incorrect or incomplete, this encompassing approach can lead to the *discovery* of a better empirical model that retains any subset of valid theory insights together with a set of variables that are substantively relevant empirically. Selection is essentially costless if not needed, and beneficial otherwise, the opposite of current beliefs.

The paper seeks to overview strategic issues influencing 'model choice' in macroeconomics. Its structure is as follows. Section 2 discusses the main criteria by which models are currently selected in macroeconomics, specifically economic theory with its subset of policy relevance, and empirical evidence with its subset of forecast accuracy. Section 3 reviews the foundations of econometrics initiated by Trygve Haavelmo, how various aspects have developed since, and the relationship between his concept of 'design of experiments' and the theory of reduction that determines the target for model selection. Section 4 considers whether empirical evidence alone can decide the choice of model, and concludes not. Section 5 highlights some major issues confronting inter-temporal macroeconomic theory: Section 5.1 considers the interlinked roles of theory and evidence, illustrated in Section 5.2 by a 'Phillips curve' example; Section 5.3 describes the consequences of unanticipated shifts of distributions, where Section 5.4 focuses on the resulting difficulties for theories of expectations formation, and Section 5.5 on the closely linked failures of the law of iterated expectations when distributions shift over time. The conclusion is that theory alone is an inadequate basis for model choice even when a theory model is the objective of an analysis, and is stringently tested, as discussed in Section 5.6. Section 6 shows that forecasting performance cannot distinguish reliably between good and bad models. Consequently, Section 7 then proposes embedding theory-driven and data-driven approaches during model selection while retaining economic theory insights. Section 7.1 explains the formulation of the initial general unrestricted model (GUM); Section 7.2 describes selecting empirical models therefrom, and outlines automatic empirical model discovery while also tackling multiple location shifts, namely sudden, often unanticipated, changes in the levels of the data processes. This provides the basis for the proposal in Section 7.3 for retaining (but not imposing) economic theory models, unaffected by selecting over many contending alternatives, while checking for location shifts of any magnitudes and signs anywhere in the sample, extended in Section 7.4 to working with an incomplete, or invalid, theory model. Section 8 discusses the implications for evaluating policy models, and Section 9 considers how the overall analysis might help clarify approaches to nowcasting and flash estimates. Section 10 concludes.

2. Criteria for deciding between alternative approaches

Many criteria have been used to select models in macroeconomics, including: theory generality, internal consistency, insights, invariance, novelty, excess content, policy relevance, identification, and consistency with evidence; empirical goodness-of-fit, congruence, constancy, parsimony, encompassing, consistency with theory, and forecast accuracy; as well as elegance, relevance, telling a story, and making money, *inter alia*. An obvious solution to resolve the dilemma as to which criteria should be used is to match them all. Unfortunately, some criteria conflict (e.g., generality versus parsimony; elegance versus congruence, etc.), human knowledge is limited, and economies

[1] See Harrison et al. (2005), to be replaced by the new dynamic stochastic general equilibrium (DSGE) model in Burgess et al. (2013), called COMPASS (Central Organising Model for Projection Analysis and Scenario Simulation). http://bankunderground.co.uk/2015/11/20/how-did-the-banks-forecasts-perform-before-during-and-after-the-crisis/ provides an honest appraisal of the failures of the Bank's new model even on past data.

are high dimensional, heterogeneous, and non-stationary, with data samples that are often small, incomplete and may be inaccurate. Maximizing goodness of fit or likelihood is fine for estimating a given specification, but is inappropriate for selecting specifications; and attempts to improve inference about parameters of interest by selection have been criticized by Leeb and Pötscher (2003, 2005). Trying to improve forecasting by selection or theory imposition has not prevented intermittent forecast failure.

Since the potential uses of an empirical model may range across data description, testing theories, policy analyses and forecasting, models also need to be robust to: (a) mis-specification, (b) past outliers and shifts, (c) regime changes, and (d) aggregation over space, time and variables. The first requires either orthogonal omissions or a correct and comprehensive economic theory combined with a complete choice of all relevant variables, lags and functional forms. The second necessitates modelling all forms of breaks, some of which may have arisen from past policy changes themselves, so would question the reliability of models to (c). The fourth entails finding similar data descriptions at different aggregation levels. As unanticipated shifts of many magnitudes occur intermittently, (e) forecasts need to be robust to recent shifts, even if forecasting shifts remains an elusive goal. Whether achieving all five requirements jointly is even potentially feasible is unknown, but it seems unlikely with most extant approaches, although all five are testable, at least in part. Below we describe an approach that retains useful theoretical insights, without imposing the theory, when selecting from large numbers of variables, functional forms and shift indicators, with tests of invariance to tackle (a)–(c), as in Doornik and Hendry (2015).

In an overly simple summary, the description of the present state of macroeconomics sketched in the introduction also seems to have been the prevailing situation when Trygve Haavelmo (1944) wrote his *Probability Approach*. Haavelmo initially assumed that relevant and valid macroeconomic theory existed, but viable econometric methods to quantify it did not, so he developed a general framework for the latter. His foundations for econometrics set the scene for the vast and powerful advances that have followed since: see Hendry and Morgan (1995), Morgan (1990) and Qin (1993, 2013) for histories of econometrics. However, Haavelmo soon found that there was in fact little empirically-relevant economic theory, so he commenced research on economics, as described, for example, in his Presidential Address to the Econometric Society (see Anundsen, Nymoen, Krogh, and Vislie, 2012, and Haavelmo, 1958) and later in his Nobel Lecture (see Haavelmo, 1989): Bjerkholt (2005) provides valuable discussion.

Haavelmo (1944) both formalized analytical methods for econometrics in a general probabilistic framework, and clarified a range of specific stochastic models and their properties. A key aspect was his notion of a *design of experiments* to match the data generation process, so valid inferences could be conducted. Section 3 addresses how the theory of reduction (see Hendry, 1987) helps formalize that idea. 'Selection' in Haavelmo (1944) refers to choosing the theory model together with the specification of any functions, variables and parameters, whereas 'testing' appears to include the notion of selection as used below, namely checking how well a theory model matches the evidence, as well as the usual sense of both specification and mis-specification testing: see Spanos (1989) and Juselius (1993). In a chapter added after the 1941 version (see Bjerkholt, 2007), Haavelmo (1944) also formalized the conditions for successful economic forecasting, requiring that the model in use was correct and constant—highlighting the key role of constancy of the joint distribution between the sample period and the forecast horizon—but he seems to have been unaware of the earlier analysis in Smith (1929). We revisit forecasting as a device for model choice in Section 6.

Major changes have occurred in the conceptualization of economic time series since Haavelmo wrote, in particular their non-stationarity. Economies change continually and sometimes abruptly, forcing a careful consideration of stochastic trends from unit roots in the dynamic representation and structural shifts in DGPs. While these features complicate empirical analyses, and developing appropriate tools for valid inferences in such settings has been a major—and remarkably successful—enterprise, less attention has been paid to their pernicious implications for macroeconomic theory, economic policy analyses and forecasting, an important theme below. A second key development has been in viable methods of selecting empirical models despite all the complications of aggregate economic data and the need to allow for more variables (N) than observations (T) in the candidate set of determinants to be considered.

Many features of empirical models are not derivable from abstract theory, so must be based on the available data sample to discover what actually matters empirically. Since *ceteris paribus* is usually inappropriate in macroeconomics, specifications must include the complete set of variables that are actually relevant, their lagged responses and functional forms, non-stationarities such as structural breaks and unit roots, and appropriate exogeneity conditions. Consequently, while economic theory may provide the *object* of interest within modelling, it cannot be the *target* of an empirical study, although it is often imposed as the latter. The next section considers what the relevant target might be, and how it matches Haavelmo's concept of 'design of experiments'.

3. 'Design of experiments' and the target for model selection

As Haavelmo (1944, p. 14) expressed the matter: "We try to choose a theory and a design of experiments to go with it, in such a way that the resulting data would be those which we get by passive observation of reality". At first sight, it is not obvious how we can choose a 'design of experiments' in macroeconomics to match the aggregate observed data, but the theory of reduction provides a possible interpretation.

To understand how an economy functions, the appropriate target must be its DGP. That DGP is the joint density of all the variables in the economy under analysis, so is too high dimensional and non-stationary to develop

complete theories about, or to model empirically. The local DGP (denoted LDGP) is the DGP for the r variables $\{\mathbf{w}_t\}$ which an investigator has chosen to model over $t = 1, \ldots, T$, usually based on a prior subject-matter theory. The theory of reduction explains the derivation of the LDGP from the DGP, leading to the joint density $\mathsf{D}_{\mathbf{W}_T^1}(\mathbf{w}_1 \ldots \mathbf{w}_T | \mathbf{Q}_T^1, \boldsymbol{\theta}_T^1, \mathbf{W}_0)$ where $\mathbf{W}_T^1$ denotes the complete set of observed data $(\mathbf{w}_1 \ldots \mathbf{w}_T)$, $\mathbf{Q}_T^1 = (\mathbf{q}_1 \ldots \mathbf{q}_T)$ are all the deterministic terms, and the entailed 'parameters' $\boldsymbol{\theta}_T^1$ of the economic agents' decision processes may be time varying (see e.g., Hendry, 2009). Following Doob (1953), by sequential factorization, the LDGP $\mathsf{D}_{\mathbf{W}_T^1}(\cdot)$ can always be written with a martingale difference error, or innovation, that is unpredictable from the past of the process:

$$\begin{aligned} \mathsf{D}_{\mathbf{W}_T^1}\left(\mathbf{W}_T^1 \mid \mathbf{Q}_T^1, \boldsymbol{\theta}_T^1, \mathbf{W}_0\right) &= \mathsf{D}_{\mathbf{w}_T}\left(\mathbf{w}_T | \mathbf{W}_{T-1}^1, \mathbf{q}_T, \boldsymbol{\theta}_T, \mathbf{W}_0\right) \\ &\quad \times \mathsf{D}_{\mathbf{W}_{T-1}^1}\left(\mathbf{W}_{T-1}^1 | \mathbf{Q}_{T-1}^1, \boldsymbol{\theta}_{T-1}^1, \mathbf{W}_0\right) \\ &\quad \vdots \\ &= \prod_{t=1}^{T} \mathsf{D}_{\mathbf{w}_t}\left(\mathbf{w}_t \mid \mathbf{W}_{t-1}^1, \mathbf{q}_t, \boldsymbol{\theta}_t, \mathbf{W}_0\right). \end{aligned} \tag{1}$$

Thus, the joint density can be expressed as the product of the individual conditional densities. Let $\boldsymbol{\epsilon}_t = \mathbf{w}_t - \mathsf{E}_{\mathsf{D}_{\mathbf{w}_{t-1}}}\left[\mathbf{w}_t | \mathbf{W}_{t-1}^1, \mathbf{q}_t\right]$, where the expectation operator must be subscripted by the specific distribution over which the integral is calculated (see Section 5.4), then $\mathsf{E}_{\mathsf{D}_{\mathbf{w}_{t-1}}}\left[\boldsymbol{\epsilon}_t | \mathbf{W}_{t-1}^1, \mathbf{q}_t\right] = \mathbf{0}$, so that $\{\boldsymbol{\epsilon}_t\}$ is an innovation error process relative to the available information, providing a basis for laws of large numbers and central limit theorems.

Consequently, the LDGP provides an appropriate 'design of experiments' whereby the passively observed data are described to within the smallest possible innovation errors given the choice of $\{\mathbf{w}_t\}$. The LDGP innovation error $\{\boldsymbol{\epsilon}_t\}$ is **designed**, or created, by the reductions entailed in moving from the DGP to the distribution of $\{\mathbf{w}_t\}$, so is not an 'autonomous' process. A 'better' choice of variables than $\{\mathbf{w}_t\}$, namely one where the LDGP is closer to the actual DGP, would deliver yet smaller innovation errors. Nevertheless, once $\{\mathbf{w}_t\}$ is chosen, one cannot do better than know $\mathsf{D}_{\mathbf{W}_T^1}(\cdot)$, which encompasses all models thereof on the same data or subsets thereof (see Bontemps and Mizon, 2008). Given $\{\mathbf{w}_t\}$, the LDGP is therefore the appropriate *target* for model selection. The set of variables $\{\mathbf{w}_t\}$ chosen for analysis will invariably depend on the subject-matter theory, institutional knowledge, and previous evidence, so in that sense any theory-model *object* is always directly related to the target LDGP.

However, while the LDGP provides the appropriate 'design of experiments', it is unknown in practice, so must be discovered from the available evidence. Consequently, (1) really provides cold comfort for empirical modelling: the sequential factorization delivers an innovation error only using the correct $\mathsf{D}_{\mathbf{w}_t}(\cdot)$ at each point in time, including knowledge of the relevant $\{\mathbf{q}_t\}$, which is why Hendry and Doornik (2014) emphasize the need to discover the LDGP. Doing so involves nesting that LDGP in a suitably general unrestricted model (denoted GUM) while also embedding the theory model in that GUM, then searching for the simplest acceptable representation thereof, and stringently evaluating that selection for congruence and encompassing. Section 7 addresses that process: first we consider whether in isolation empirical evidence, theory, or forecasting can decide model choice in macroeconomics.

4. Can empirical evidence alone decide?

Two key aspects of the complexity of macroeconomics are its high dimensionality and its non-stationarity from stochastic trends and the occurrence of numerous unanticipated location shifts, altering the underlying data distributions. The recent financial crisis is just the latest example of such shifts, which are why all the densities and expectations in the previous section were subscripted by their time-dated variables' distributions. Concerning the first, and notwithstanding recent major advances in modelling 'big data', it is impossible to proceed in empirical economics without some *a priori* ideas of the key influences to include in a model. Empirical evidence alone cannot decide.

Location shifts are important in Section 5 in explaining the 'break-down' of empirical macroeconomic models, and in Section 6 in accounting for forecast failures. Conversely, as location shifts are rarely included in macroeconomic theories, empirical evidence remains essential, so is necessary but not sufficient. Numerous historical examples of such shifts are discussed in Hendry (2015), and as a further illustration, Fig. 1a shows year-on-year percentage changes in Japanese exports, which look 'well behaved' over 2000–2008, then drop by **70**% in mid 2008–2009, as shown in panel b. Such a dramatic fall was not anticipated at the time, nor were the subsequent gyrations. Many of the shifts found for (e.g.) annual changes in UK real GDP correspond to major policy changes, and most were also not anticipated at the time. Figs. 1 and 2 (shown below) emphasize that growth rates need not be stationary processes: differencing a data series certainly removes a unit root, but cannot ensure stationarity. However, outliers and shifts can be accommodated in empirical modelling by indicator saturation methods, either impulses (IIS: see Hendry, Johansen, and Santos, 2008, and Johansen and Nielsen, 2009, 2016), or steps (SIS: see Castle, Doornik, Hendry, and Pretis, 2015: Ericsson and Reisman, 2012, propose combining these in 'super saturation'): all are already implemented in *Autometrics* (see Pretis, Reade, and Sucarrat, in press, for an R version) as the next section illustrates.

5. Is theoretical analysis definitive?

We first consider the relationship between theory and evidence in Section 5.1, followed by a 'Phillips curve' illustration in Section 5.2, then describe the serious problems confronting inter-temporal macroeconomic theory facing shifting distributions in Section 5.3, focusing on the difficulties for theories of expectations formation in Section 5.4 and the law of iterated expectations in Section 5.5.

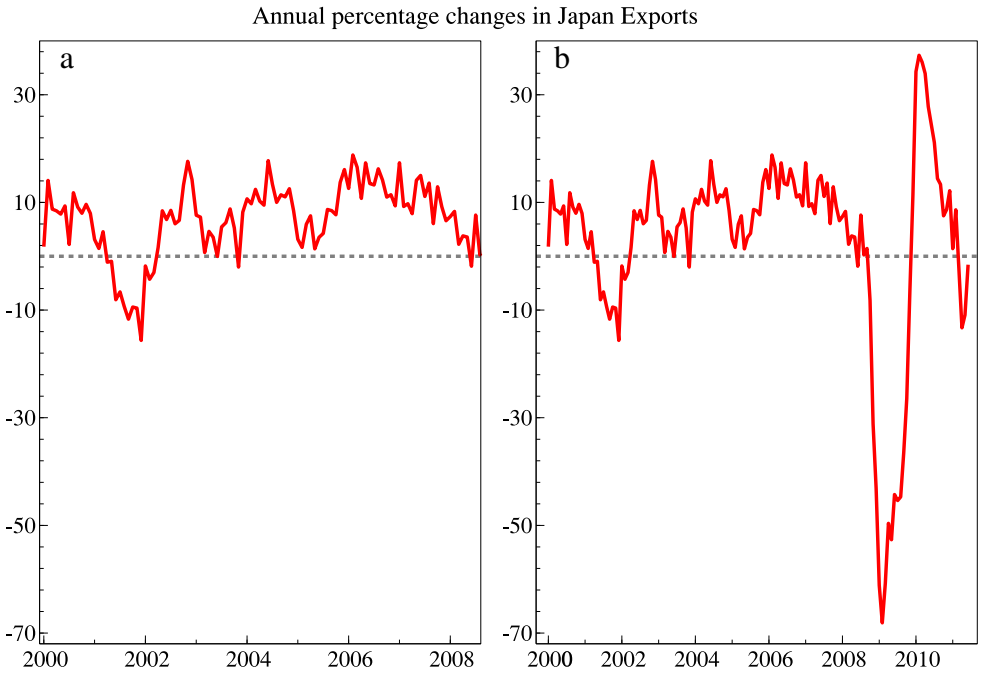

Fig. 1. Year-on-year percentage changes in Japanese exports.

5.1. Theory and evidence

Past developments and discoveries in economics have primarily come from abstract theoretical reasoning. However, all economic theories are by their nature incomplete, incorrect, and mutable, needing strong *ceteris paribus* claims, which are inappropriate in a non-stationary, evolving world (see Boumans and Morgan, 2001), and relying on many relatively arbitrary assumptions. Say a macroeconomic analysis suggests a theoretical relationship of the form:

$$\mathbf{y}_t = \mathbf{f}(\mathbf{x}_t) \tag{2}$$

where the vector of k variables denoted $\mathbf{y}_t$ depend on n 'explanatory' variables $\mathbf{x}_t$, which may include lagged values and expectations. There are $m > n$ instrumental variables $\mathbf{z}_t$, believed to be 'exogenous' and to influence $\mathbf{x}$.

However, the form of $\mathbf{f}(\cdot)$ in (2) depends on the utility or loss functions of the relevant agents, the constraints they face, the information they possess and how they form their expectations. To become operational, analyses need to assume some functional forms for $\mathbf{f}(\cdot)$, that $\mathbf{f}(\cdot)$ is constant, that only $\mathbf{x}$ matters, what a time unit measures (days or decades), and that the $\{\mathbf{z}_t\}$ really are 'exogenous' even though available data will have been aggregated over space, commodities, time and heterogeneous individuals whose endowments can shift, often abruptly, as in the 2008 financial crisis. Moreover, economic theory evolves, improving our understanding of the world and often changing that world itself, from the introduction of the concept of the 'invisible hand' in Adam Smith (1759), through key insights into option pricing, auctions and contracts, principal–agent and game theories, trust, moral hazard, and asymmetric information *inter alia*. Consequently imposing contemporaneous economic-theory models in empirical research entails that applied findings will be forgotten when the economic theory that they quantified is discarded. This is part of the explanation for the fads and fashions, cycles and schools in economics.

5.2. Inflation, unemployment and 'Phillips curves'

Location shifts in individual time series will induce breaks in relationships unless the shifts coincide across variables, called co-breaking (see Hendry and Massmann, 2007). A 'classic' example of the failure of supposedly causally-connected variables to co-break is the 'Phillips curve' linking wage inflation, Δw, and unemployment, U_r. The long-run history of their many location shifts is shown in Fig. 2 panels *a* and *c*. The shifts portrayed are selected by SIS at 0.1%, and are numbered (1)–(10) for wage inflation and [1]–[11] for unemployment. Table 1 explains the interpretation of the numbered shifts, all of which correspond to key events in UK economic and political history.

Although many of the major events are in common, their timing does not necessarily overlap, and the right-hand two panels in Fig. 2, *b* and *d*, respectively show their mismatched location shifts over time, and the consequent shifts in the 'Phillips curves' in different epochs. Even excluding war years, price indexation periods, price controls, and the aftermath of World War I, the curves shift. Such a finding does not entail that there is no constant relationship between wage inflation and unemployment, merely that a bivariate 'Phillips curve' is not such a relation. Castle and Hendry (2014b) present a model relating real wages to a number of variables including unemployment, where almost all the shifts seen in Fig. 2 then co-break: it

Table 1
Major shifts in annual UK wage inflation, Δw_t and unemployment, $U_{r,t}$.

Δw	Event	U_r	Event
(1)	UK 'Great Depression'	[1]	Business cycle era
(2)	World War I	[2]	UK 'Great Depression'
(3)	Postwar crash	[3]	World War I
(4)	US 'Great Depression'	[4]	Postwar crash
(5)	World War II	[5]	US crash
(6)	Postwar reconstruction	[6]	Leave Gold Standard
(7)	'Barber boom'	[7]	World War II
(8)	Oil crisis	[8]	Postwar reconstruction
(9)	Leave ERM	[9]	Oil crisis and Mrs Thatcher
(10)	'Great Recession'	[10]	Leave ERM
		[11]	'Great Recession'

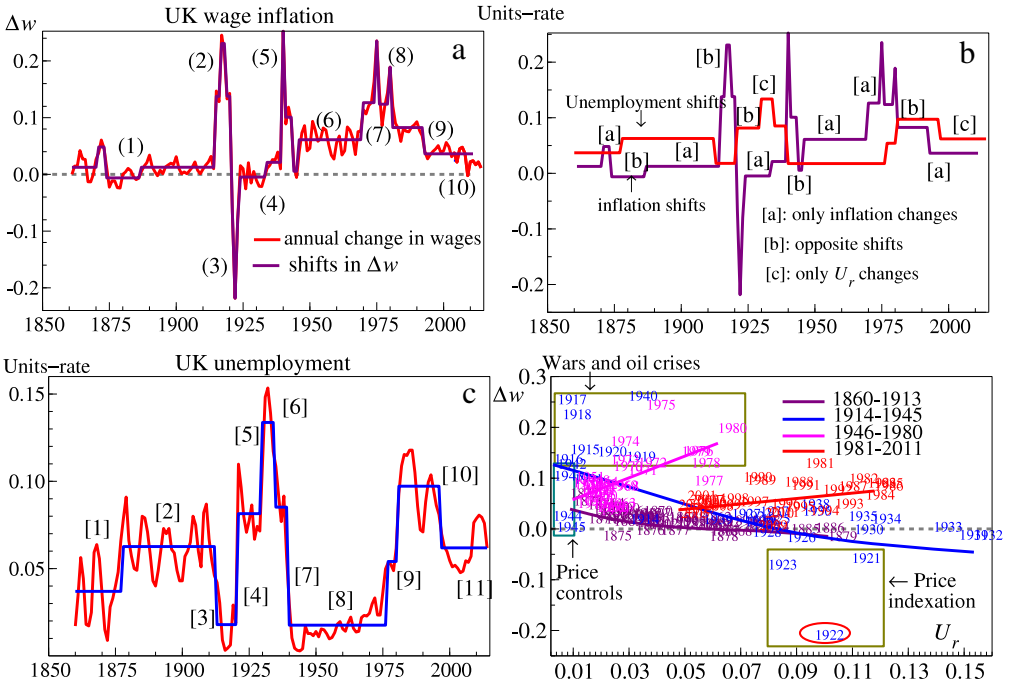

Fig. 2. Distinct epochs and major shifts in annual UK wage changes, Δw_t, and unemployment, $U_{r,t}$.

is essential to model shifts to understand the underlying economic behaviour.

5.3. Problems confronting inter-temporal macroeconomic theory

The three aspects of unpredictability delineated in Hendry and Mizon (2014) help explain the intermittent failure evident in Fig. 2. *Intrinsic unpredictability* occurs in a known distribution, arising from chance distribution sampling, 'random errors', etc., so is essentially an unknown known. Empirically, population distributions are never known, but nevertheless much of statistical theory depends on random sampling from distributions with assumed known properties. *Instance unpredictability* arises in (for example) fat-tailed distributions from outliers of unknown magnitudes and signs at unanticipated times (known unknowns, aka 'black swans' from Taleb, 2007). *Extrinsic unpredictability* (unknown unknowns) occurs from unanticipated shifts of distributions from unknown numbers, signs, magnitudes and timings of shifts (as in the concept of reflexivity in Soros, 2008). Each type of unpredictability has different effects on economic theory, policy analyses, forecasting and nowcasting, as well as econometric modelling. The first three can go awry from instance or extrinsic unpredictability as explained in the next two sections, yet empirical outcomes are susceptible to being modelled *ex post*, as we explain below.

5.4. Problems with expectations formation

Let f_{x_t} denote the density of a random variable x_t at time t, and writing the derivation as is often the way in economics, let $x_{t+1} = \mathsf{E}[x_{t+1}|\mathcal{I}_t] + \nu_{t+1}$ where $\mathsf{E}[\cdot]$ is the assumed expectations operator and $\mathcal{I}_t$ is the information available at t. Then $\mathsf{E}[\nu_{t+1}|\mathcal{I}_t] = 0$. This derivation may be mistakenly thought to establish an unbiased expectation of the future value x_{t+1}. However, absent a crystal

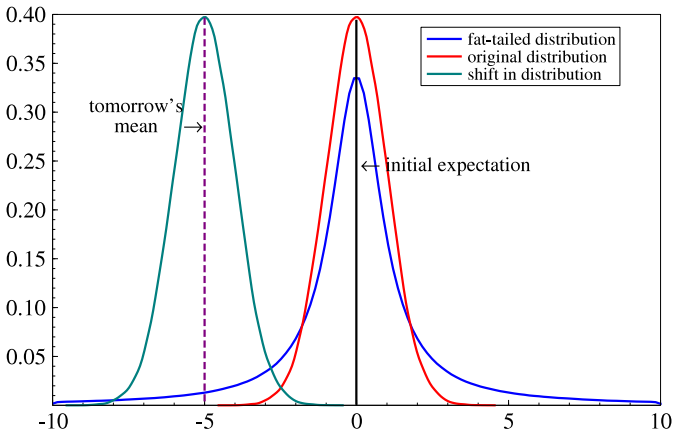

Fig. 3. Location shifts of distributions.

ball, the derivation merely shows that $\mathsf{E}_{\mathsf{f}_{x_t}}[\nu_{t+1}|\mathcal{I}_t] = 0$. Nothing can ensure that when the next period eventuates, $\mathsf{E}_{\mathsf{f}_{x_{t+1}}}[\nu_{t+1}|\mathcal{I}_t] = 0$, as an unbiased expectation of the future value x_{t+1} requires, since:

$$\mathsf{E}_{\mathsf{f}_{x_{t+1}}}[x_{t+1} \mid \mathcal{I}_t] = \int x_{t+1}\mathsf{f}_{x_{t+1}}(x_{t+1} \mid \mathcal{I}_t)\,\mathrm{d}x_{t+1} \tag{3}$$

Unfortunately, (3) requires knowing the entire *future* distribution $\mathsf{f}_{x_{t+1}}(x_{t+1}|\mathcal{I}_t)$. The best an economic agent can do is form a sensible expectation, forecasting $\mathsf{f}_{x_{t+1}}(\cdot)$ by $\widehat{\mathsf{f}}_{x_{t+1}}(\cdot)$. But, if the moments of $\mathsf{f}_{x_{t+1}}(\cdot)$ alter, there are no good rules for $\widehat{\mathsf{f}}_{x_{t+1}}(\cdot)$, although $\widehat{\mathsf{f}}_{x_{t+1}}(\cdot) = \mathsf{f}_{x_t}(\cdot)$ is **not** usually a good choice facing location shifts. Agents cannot know $\mathsf{f}_{x_{t+1}}(\cdot)$ when there is no time invariance, nor how $\mathcal{I}_t$ will enter the conditional expectation at time $t+1$. Deducing that agents can make unbiased predictions of future outcomes by writing $\mathsf{E}[\nu_{t+1}|\mathcal{I}_t] = 0$ is sleight of hand, not subscripting the expectation by the relevant distribution, as in (e.g.) (1).

Fig. 3 illustrates the potential consequences of distributional shifts. The right-most densities are a standard normal with a fat-tailed (t_3) density superimposed, often leading to the occurrence of a 'black swan' event, namely an unlikely draw from a known distribution. However, an unanticipated location shift as in the left-most density will lead to a 'flock' of black swans, which would be most unlikely for independent draws from the original distribution. Moreover, after such a location shift, the previous conditional mean can be a very poor representation of the new one as shown. This problem also impacts on using the law of iterated expectations to derive 'optimal' decision rules as follows.

5.5. Failures of the law of iterated expectations

When variables at different dates, say (x_t, x_{t+1}), are drawn from the **same** distribution f_x, then:

$$\mathsf{E}_{\mathsf{f}_x}\left[\mathsf{E}_{\mathsf{f}_x}[x_{t+1} \mid x_t]\right] = \mathsf{E}_{\mathsf{f}_x}[x_{t+1}] \tag{4}$$

In terms of integrals:

$$\begin{aligned}\mathsf{E}_{\mathsf{f}_x}\left[\mathsf{E}_{\mathsf{f}_x}[x_{t+1} \mid x_t]\right] &= \int_{x_t}\left(\int_{x_{t+1}} x_{t+1}\mathsf{f}_x(x_{t+1}|x_t)\,\mathrm{d}x_{t+1}\right)\mathsf{f}_x(x_t)\,\mathrm{d}x_t \\ &= \int_{x_{t+1}} x_{t+1}\left(\int_{x_t}\mathsf{f}_x(x_{t+1}, x_t)\,\mathrm{d}x_t\right)\mathrm{d}x_{t+1} \\ &= \int_{x_{t+1}} x_{t+1}\mathsf{f}_x(x_{t+1})\,\mathrm{d}x_{t+1} \\ &= \mathsf{E}_{\mathsf{f}_x}[x_{t+1}]\end{aligned} \tag{5}$$

confirming that the law of iterated expectations holds inter-temporally for constant f_x.

However, when the distribution shifts between t and $t+1$, so $\mathsf{f}_{x_t}(x_t) \neq \mathsf{f}_{x_{t+1}}(x_t)$, then:

$$\mathsf{E}_{\mathsf{f}_{x_t}}\left[\mathsf{E}_{\mathsf{f}_{x_t}}[x_{t+1} \mid x_t]\right] \neq \mathsf{E}_{\mathsf{f}_{x_{t+1}}}[x_{t+1}]\,. \tag{6}$$

Hendry and Mizon (2014) provide formal derivations. Unanticipated shifts of distributions between time periods–as occur intermittently–invalidate inter-temporal derivations. Thus, economic theories based on assuming (4) holds require that no breaks ever occur—so are empirically irrelevant—or will always suffer structural breaks when distributions shift, as the mathematics behind their derivations fails when such shifts occur. This fundamental difficulty is not merely a problem for empirical models based on DSGE theory, it is just as great a problem for the economic agents whose behaviour such models claim to explain. Consequently, DSGEs suffer a double flaw: they are making incorrect assumptions about the behaviour of the agents in their model, and are also deriving false implications therefrom by using mathematics that is invalid when applied to realistically non-stationary economies. Other criteria than theoretical elegance, 'constant-distribution rationality', or generality are required to decide between alternative approaches.

5.6. Problems with testing pre-specified models

As will be shown in the next section, forecast performance is not a reliable guide to the verisimilitude of an empirical model, so theory-based, or more generally, pre-specified, models are often evaluated by testing various of their assumptions and/or implications. Even when the null rejection probability of multiple tests is appropriately controlled, there are two fundamental problems with such a strategy. First, tests can fail to reject over a wider range

of states of nature than that conceived as the 'null hypothesis', namely their implicit null hypothesis as discussed by Mizon and Richard (1986); and can reject for completely different reasons when their null is true, such as a test for error autocorrelation rejecting when autocorrelated residuals are created by an unmodelled location shift. Consequently, it is a non-sequitur to infer the alternative holds when the null is rejected. Second, when any one of the tests being calculated rejects, whether or not any others do is uninformative as to their validity. Worse, 'fixing' a sequence of rejections by adopting their erstwhile alternative compounds these problems as an untested misspecification may have induced all the rejections. At its heart, the fundamental difficulty with this strategy is going from the simple to the general, so the viability of each test-based decision is contingent on the 'correctness' of *later* test outcomes. Section 7 proposes our alternative approach.

6. Can forecasting performance help distinguish good from bad models?

Unanticipated location shifts also have adverse consequences both for forecast accuracy, and for the possibility of using forecast outcomes to select between models. As shown in Clements and Hendry (1998), while 'good' models of a DGP can forecast well, and 'poor' models can forecast badly, it is also true that 'good' models, including the in-sample DGP, can forecast badly and 'poor' models can forecast well. A well-worn analogy (based on Apollo 13) is of a rocket that is forecast to land on the moon on 4th July of a given year, but en route is hit by a meteor and knocked off course, so never arrives at its intended destination. Such a forecast is badly wrong, but the failure is neither due to a poor forecasting model nor does it refute the underlying Newtonian gravitation theory. In terms of this paper, the failure is due to an unanticipated distribution shift: there is no theorem proving that the DGP up to time T is always the best device for forecasting the outcome at $T+1$ in a wide-sense non-stationary process, especially so if the parameters of that DGP need to be estimated from the available data evidence. Conversely, some methods are relatively robust after location shifts so insure against *systematic* forecast failure: producing the (relatively) best forecasting outcome need not entail a good model in terms of verisimilitude.

To illustrate, consider an in-sample stationary conditional DGP given by the equilibrium-correction mechanism (EqCM) over $t = 1, \ldots, T$:

$$\begin{aligned}\Delta x_t &= \mu + (\rho - 1)\, x_{t-1} + \gamma z_t + \epsilon_t \\ &= (\rho - 1)\,(x_{t-1} - \theta) + \gamma(z_t - \kappa) + \epsilon_t \end{aligned} \tag{7}$$

where $|\rho| < 1$, $\epsilon_t \sim \mathsf{IN}[0, \sigma_\epsilon^2]$, the past, present and future values of the super-exogenous variable $\{z_t\}$ are known, with $\mathsf{E}[z_t] = \kappa$ where $\mathsf{E}[x_t] = \theta = (\mu + \gamma\kappa)/(1 - \rho)$. Providing (7) remains the DGP of $\{x_t\}$ for $t > T$, forecasts based on that as a model will perform as anticipated. However, problems arise when the parameters shift, as in:

$$\begin{aligned}\Delta x_{T+h} = \left(\rho^* - 1\right)\left(x_{T+h-1} - \theta^*\right) + \gamma^*(z_{T+h} - \kappa^*) \\ + \epsilon_{T+h}\end{aligned} \tag{8}$$

where at time T, every parameter has changed but with $|\rho^*| < 1$ (we could allow σ_ϵ^2 to change as well). The key problem for forecasting transpires to be shifts in the long-run mean from the in-sample value $\mathsf{E}_{\mathsf{f}_{x_t}}[x_t] = (\mu + \gamma\kappa)/(1 - \rho) = \theta$ for $t \leq T$ to $\mathsf{E}_{\mathsf{f}_{x_{T+h}}}[x_{T+h}] = (\mu^* + \gamma^*\kappa^*)/(1 - \rho^*) = \theta^*$ for $t \gg T$. Forecast failure is unlikely if $\theta = \theta^*$, but almost certain otherwise: see Hendry and Mizon (2012).

All EqCMs fail systematically when θ changes to θ^*, as their forecasts automatically converge back to θ, irrespective of the new parameter values in (8). EqCMs are a huge class including most regressions; dynamic systems; vector autoregressions (VARs); cointegrated VARs (CVARs); and DSGEs; as well as volatility models like autoregressive conditional heteroskedasticity (ARCH), and its generalization to GARCH. Systematic forecast failure is a pervasive and pernicious problem affecting all EqCM members.

6.1. Forecast-error sources

When (7) is the in-sample DGP and (8) the DGP over a forecast horizon from T, consider an investigator who tries to forecast using an estimated first-order autoregression, often found to be one of the more successful forecasting devices (see e.g., Stock and Watson, 1999):

$$\widehat{x}_{T+1|T} = \widehat{\theta} + \widehat{\rho}(x_T - \widehat{\theta}) \tag{9}$$

where $\widehat{\theta}$ estimates the mean of the process, from which the intercept, $\widehat{\mu} = (1 - \widehat{\rho})\widehat{\theta}$, can be derived. Almost every possible forecast error occurs using (9) when (8) is the forecast-horizon DGP, namely:

$$\begin{aligned}\widehat{\epsilon}_{T+1|T} = \left(\theta^* - \widehat{\theta}\right) + \rho^*\left(x_T - \theta^*\right) - \widehat{\rho}\left(\widehat{x}_T - \widehat{\theta}\right) \\ + \gamma^*\left(z_{T+1} - \kappa^*\right) + \epsilon_{T+1}\end{aligned} \tag{10}$$

(ia) deterministic shifts: $(\theta,\ \kappa)$ changes to $(\theta^*,\ \kappa^*)$;
(ib) stochastic breaks: $(\rho,\ \gamma)$ changes to $(\rho^*,\ \gamma^*)$;
(iia,b) inconsistent parameter estimates: $\theta_e \neq \theta$ and $\rho_e \neq \rho$ where $\rho_e = \mathsf{E}[\widehat{\rho}]$ and $\theta_e = \mathsf{E}[\widehat{\theta}]$;
(iii) forecast origin uncertainty: $\widehat{x}_T \neq x_T$;
(iva,b) estimation uncertainty: $\mathsf{V}[\widehat{\rho},\ \widehat{\theta}]$;
(v) omitted variables: z_{T+1} was inadvertently excluded from (9);
(vi) innovation errors: ϵ_{T+1}.

Clearly, (9) is a hazardous basis for forecasting, however well it performed as an in-sample model.

Indeed even if all in-sample breaks have been handled, and there are good forecast origin estimates so $\widehat{x}_T = x_T$, then taking expectations in (10) and noting that $\mathsf{E}_{x_T, z_{T+1}}[x_T - \theta] = 0$, $\mathsf{E}_{x_T, z_{T+1}}[z_{T+1} - \kappa^*] = 0$, and $\mathsf{E}_{x_T, z_{T+1}}[\epsilon_{T+1}] = 0$:

$$\mathsf{E}_{x_T, z_{T+1}}[\widehat{\epsilon}_{T+1|T}] \simeq \left(1 - \rho^*\right)\left(\theta^* - \theta\right) \tag{11}$$

irrespective of model mis-specification. The forecast bias on the right-hand side of (11) persists so long as (9) continues to be used, even though no further breaks ensue. Crucially, there is no systematic bias when $\theta^* = \theta$: the

values of μ and ρ can change by large magnitudes provided $\theta = \theta^*$, as the outcome is isomorphic to $\mu = \mu^* = 0$: see Clements and Hendry (1999). In that setting, there would be almost no visible evidence from which agents, including policy makers, could ascertain that the DGP had changed, although later policy effects could be radically different from what was anticipated.

Surprisingly, even if $\mu = \mu^* = 0$ and the variable z_t is correctly included in the forecasting model, but itself needs to be forecast, then $\rho \neq \rho^*$ still induces forecast failure by shifting $\theta \neq \theta^*$ when $\kappa \neq 0$. Thus, the only benefit from including z_t in avoiding forecast failure is if κ alone shifts to κ^* which induces a shift in θ to θ^* that is captured. Nevertheless, the in-sample DGP need not forecast well.

Conversely, there is a way to forecast after location shifts that is more robust, as described in (e.g.) Castle, Clements, and Hendry (2015). Difference the mis-specified model (9) after estimation:

$$\Delta\widetilde{x}_{T+h|T+h-1} = \widehat{\rho}\,\Delta x_{T+h-1} \tag{12}$$

which entails forecasting by:

$$\begin{aligned}\widetilde{x}_{T+h|T+h-1} &= x_{T+h-1} + \widehat{\rho}\,(x_{T+h-1} - x_{T+h-2}) \\ &= \widetilde{\theta}^* + \widehat{\rho}\left(x_{T+h-1} - \overline{\theta}^*\right)\end{aligned} \tag{13}$$

Despite being incorrectly differenced, using the 'wrong' $\widehat{\rho} \neq \rho^*$, and omitting the relevant variable z_{T+h}, (13) can avoid systematic forecast failure once $h > 2$, as $\widetilde{\theta}^* = x_{T+h-1}$ and $\overline{\theta}^* = x_{T+h-2}$ are respectively 'instantaneous estimators' of θ^* in its two different roles in (13). As Castle et al. (2015) discuss, interpreted generally, (13) provides a new class of forecasting devices, with 'instantaneous estimators' at one extreme, making a highly adaptive but high variance device, and the full-sample estimator $\widehat{\theta}$ at the other, which is fixed irrespective of changes to the DGP.

The robust method works for $h > 2$ periods after the break because Δx_{T+h-1} in (12) is then:

$$\begin{aligned}\Delta x_{T+h-1} = &\left[\left(\rho^* - 1\right)\left(x_{T+h-2} - \theta^*\right) + \gamma^*(z_{T+h-1} - \kappa^*)\right] \\ &+ \epsilon_{T+h-1}.\end{aligned} \tag{14}$$

The term in brackets, $[\cdots]$, contains everything you always wanted to know when forecasting and you don't even need to ask, as Δx_{T+h-1} in (14) entails that (12) includes:
[a] the correct θ^* in $(x_{T+h-2} - \theta^*)$, so adjusts to the new equilibrium;
[b] the correct adjustment speed $(\rho^* - 1)$;
[c] z_{T+h-1} despite its omission from the forecasting device (9);
[d] the correct parameters γ^* and κ^*, with no need to estimate;
[e] the well-determined estimate $\widehat{\rho}$, albeit that ρ has shifted to ρ^*.

These features help explain the 'success' of random-walk type forecasts after location shifts, as that is the special case of (13) where only x_{T+h-1} is used, and also why updating estimates of ρ help, as tracking the location shift will drive $\widehat{\rho}$ to unity. Similar considerations apply to other adaptive forecasting devices.

To illustrate this analysis empirically, we return to 'forecasting' the outcomes for year-on-year changes in Japanese exports across 2008–2011 in Fig. 1b from models estimated using only the data in Fig. 1a. Fig. 4 records the 1-step forecasts and squared forecast errors from an autoregressive model, $\widehat{\epsilon}$, and its differenced equivalent, $\widetilde{\epsilon}$. The former has a mean-square forecast error (MSFE) of 0.56 versus 0.29 for the robust device. Consequently, it need not be informative to judge a model's verisimilitude or policy relevance by its forecast performance.

Before considering the possible role of policy relevance in selecting models, we turn to discuss the main proposal of the paper.

7. Model selection retaining economic theory insights

So far, we have established major flaws in purely empirical approaches (aka 'data driven') and purely theory-based models ('theory driven'), as well as precluded the use of forecasting success or failure as the **sole** arbiter of choice. Since the LDGP is the target, but is unknown, and a possibly incomplete theory is the object of interest within it, a joint theory-based empirical-modelling approach to discovery offers one way ahead, not simply imposing the theory model on data.

Discovery involves learning something that was previously unknown, and in the context of its time, was perhaps unknowable. As one cannot know how to discover what is not known, it is unlikely there is a 'best' way of doing so. Nevertheless, a vast number of both empirical and theoretical discoveries have been made by *homo sapiens*. Science is inductive and deductive, the former mainly reflecting the context of discovery—where 'anything goes', and an astonishing number of different approaches have occurred historically—and the latter reflects the context of evaluation—which requires rigorous attempts to refute hypotheses. Howsoever a discovery is made, it needs a warrant to establish that it is 'real', although methods of evaluation are subject-specific: economics requires a theoretical interpretation consistent with 'mainstream theory', although that theory evolves over time. Once discoveries are confirmed, accumulation and consolidation of evidence is crucial to sustain progress, inevitably requiring their incorporation into a theoretical formulation with data reduction: think $E = mc^2$ as the apex.

There are seven aspects in common to empirical and theoretical discoveries historically (see the summary in e.g., Hendry and Doornik, 2014). Discoveries always take place against a pre-existing theoretical context, or framework of ideas, yet necessitate going outside that existing state of knowledge, while searching for 'something'. That something may not be what is found, so there has to be recognition of the significance of a discovery, followed by its quantification. Next comes an evaluation of the discovery to ascertain its 'reality', and finally a parsimonious summary of the information acquired, possibly in a new theoretical framework. But there is a major drawback: experimental science is perforce simple to general, which is a slow and uncertain route to new knowledge. Globally, learning must be simple to general, but locally, it need not be. Discovery in econometrics from observational data can circumvent such limits ('big data' has implicitly noticed

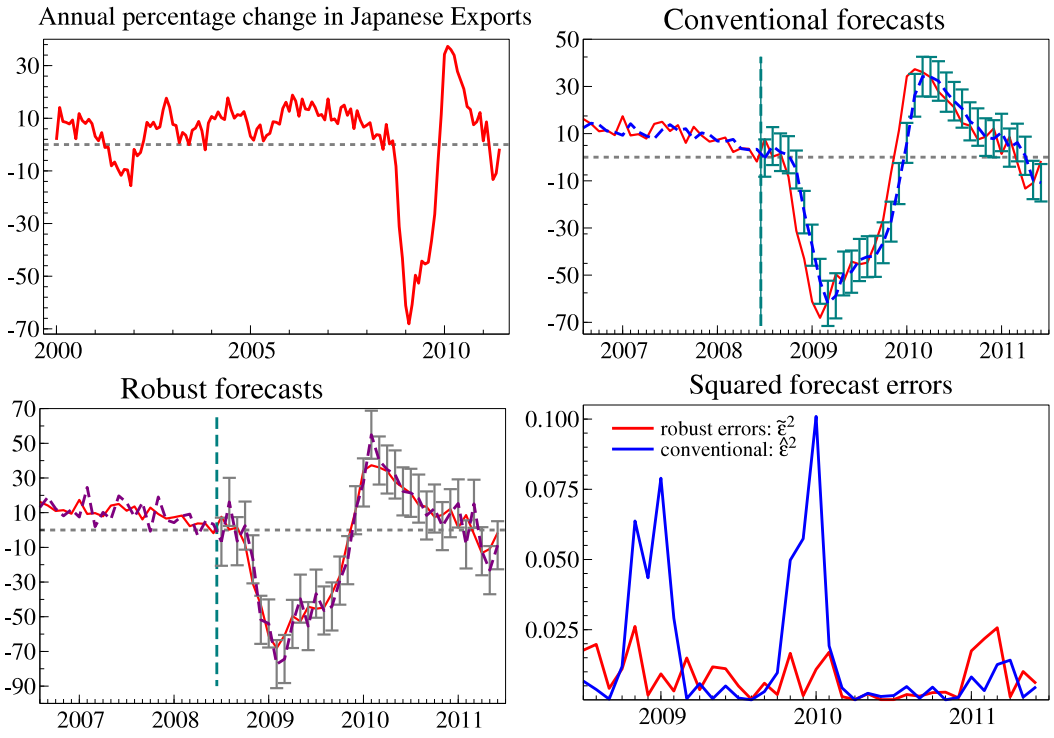

Fig. 4. 1-step 'forecasts' of Japanese exports over 2008–2011, and their squared forecast errors.

this): so how should we to proceed to capitalize on our discipline's advantage?

These seven stages have clear implications for empirical model discovery and theory evaluation in econometrics. Commencing from a theoretical derivation of a potentially relevant set of variables $\mathbf{x}$ and their putative relationships, create a more general model augmented to $(\mathbf{x}, \mathbf{v})$, which embeds $\mathbf{y} = \mathbf{f}(\mathbf{x})$. Next, automatic selection can be undertaken over orthogonalized representations, retaining the theory, to end with a congruent, parsimonious-encompassing model that quantifies the outcome in unbiasedly estimated coefficients. Because it is retained without selection, but not imposed as the sole explanation, the theory model can be evaluated directly, and any discoveries extending the context can be tested on new data or by new tests and new procedures. Finally, the possibly vast information set initially postulated can be summarized in the most parsimonious but undominated model. We now explain the detailed steps in this strategy.

7.1. Formulating the GUM

Given a theory model like (2) linking $\mathbf{y}_t$ to $\mathbf{x}_t$ which is to be maintained, extensions of it will determine how well the LDGP will be approximated. Three extensions can be created almost automatically:
(i) dynamic formulations to implement a sequential factorization as in Doob (1953);
(ii) functional form transformations for any potential non-linearity using (e.g.) the low-dimensional approach in Castle and Hendry (2010, 2014b);
(iii) indicator saturation (IIS or SIS) for outliers and parameter non-constancy: see Castle et al. (2015) and Johansen and Nielsen (2009) respectively.

These extensions are implemented jointly to create a GUM that seeks to nest the relevant LDGP and thereby deliver valid inferences. Combining the formulation decisions of the n variables $\mathbf{x}_t$, augmented by k additional candidate variables $\mathbf{v}_t$ where $r = (n+k)$, with $3r$ non-linear functions for quadratics, cubics and exponentials, all at lag length s, and any number of shifts modelled by IIS and/or SIS leads to a GUM, here formulated for a scalar y_t when the $\mathbf{u}_t$ are the principal components from $\mathbf{w}_t = (\mathbf{x}_t : \mathbf{v}_t)$ and $1_{\{i=t\}}$ are a saturating set of impulse indicators:

$$\begin{aligned} y_t = &\left[\sum_{i=1}^{n}\psi_i x_{i,t}\right] + \sum_{i=1}^{k}\phi_i v_{i,t} + \sum_{i=1}^{r}\sum_{j=1}^{s}\beta_{i,j} w_{i,t-j} \\ &+ \sum_{i=1}^{r}\sum_{j=0}^{s}\kappa_{i,j} u_{i,t-j}^2 + \sum_{i=1}^{r}\sum_{j=0}^{s}\theta_{i,j} u_{i,t-j}^3 \\ &+ \sum_{i=1}^{r}\sum_{j=0}^{s}\gamma_{i,j} u_{i,t} e^{-|u_{i,t}|} + \sum_{j=1}^{s}\lambda_j y_{t-j} \\ &+ \sum_{i=1}^{T}\delta_i 1_{\{i=t\}} + \epsilon_t. \end{aligned} \tag{15}$$

The term in brackets in (15) denotes the retained theory model: although in practice such a model may also have lags and possibly non-linearities, these are not shown for simplicity. Then (15) has $K = 4r(s+1) + s$ potential regressors, plus T indicators, so is bound to have

$N = K + T > T$ regressors in total. Variants on this formulation are possible: for example, by including separate principal components of the $\mathbf{x}_t$ and $\mathbf{v}_t$, or using different representations of shifts, such as splines, etc. The choice of the non-linear formulation is designed to include asymmetry and bounded functions, as well as approximations to such specifications as logistic smooth transition autoregressions (LSTAR: see e.g., Granger and Teräsvirta, 1993) so these can be evaluated by encompassing tests. Under the null that the theory model is complete and correct, these additions need not affect its parameter estimates given an appropriate approach with a suitably chosen nominal significance level, but under the alternative could point to a different specification as preferable. We briefly consider empirical model selection in Section 7.2 then describe how to retain theory insights in Section 7.3, and how to evaluate exogeneity in Section 8.

7.2. Selecting empirical models

Many empirical modelling decisions are undocumented as they are not recognized as involving selection. For example, a test followed by an action entails selection, as in the well-known example of testing for residual autocorrelation then using a 'correction' if the null hypothesis of white noise is rejected—albeit this practice has been criticized on many occasions (see e.g., Mizon, 1995). Unstructured selection, such as trying a sequence of variables to see if they 'matter', is rarely reported, so its statistical properties cannot be determined.

However, there are three practical ways to judge the success of **structured** selection algorithms, helping to resolve the dilemma of which evaluation criteria to use. The first is whether an algorithm can recover the LDGP starting from a nesting GUM almost as often as when commencing from the LDGP itself. Such a property entails that the costs of search are small. Next, the operating characteristics of any algorithm should match its stated properties. For example, its null retention frequency, called gauge, should be close to the claimed nominal significance level (denoted α) used in selection. Moreover, its retention frequency for relevant variables under the alternative, called potency, should preferably be close to the corresponding power for the equivalent tests on the LDGP. Note that gauge and potency differ from size and power because selection algorithms can retain variables whose estimated coefficients are insignificant on the selection statistic: see Castle, Doornik, and Hendry (2011) and Johansen and Nielsen (2016). Finally the algorithm should always find at least a well-specified, undominated model of the LDGP. All three can be checked in Monte Carlo simulations or analytically for simple models, offer viable measures of the success of selection algorithms, and can be achieved jointly as discussed more extensively in Hendry and Doornik (2014).

7.3. Retaining economic theory insights

Once $N > T$, model selection is unavoidable without incredible *a priori* claims to knowing precisely which subset of variables is relevant. Analytical evaluations across selection procedures are difficult in that setting, although operating characteristics such as gauge, potency, root mean square errors (RMSEs) after selection of parameter estimates around their DGP values, and relative success at locating the DGP, can all be investigated by simulation: Castle, Qin, and Reed (2013) provide a comprehensive comparison of the performance of many model selection approaches by Monte Carlo.

Economic theory models are often fitted directly to data to avoid possible 'model-selection biases'. This is an excellent strategy when the theory is complete and correct, but less successful otherwise. To maintain the former and avoid the latter, Hendry and Johansen (2015) propose embedding a theory model that specifies a set of n exogenous variables, $\mathbf{x}_t$, within a larger set of $n+k$ potentially relevant candidate variables, $(\mathbf{x}_t, \mathbf{v}_t)$. Orthogonalize the $\mathbf{v}_t$ to the $\mathbf{x}_t$ by projection, with residuals denoted $\mathbf{e}_t$. Then retain the $\mathbf{x}_t$ without selection, so that selection over the $\mathbf{e}_t$ by their statistical significance can be undertaken without affecting the theory parameters' estimators distributions. This strategy keeps the same theory-parameter estimates as direct fitting when the theory is correct, yet protects against the theory being under-specified when some $\mathbf{v}_t$ are relevant. Thus, the approach is an extension of encompassing, where the theory model is always retained unaltered, but simultaneously evaluated against all likely alternatives, so it merges theory-driven and data-driven approaches.

However, there are two distinct forms of under-specification. The first concerns omitting from the GUM lags, or functions, of variables in $\mathbf{w}_t$ with non-zero coefficients in the LDGP, or not checking for substantively relevant shifts, both of which can be avoided by formulating a sufficiently general initial model. The second occurs on omitting variables, say $\boldsymbol{\eta}_t$, from $\mathbf{w}_t = (\mathbf{x}_t, \mathbf{v}_t)$ where the $\boldsymbol{\eta}_t$ are relevant in the DGP, which induces a less useful LDGP (one further from the actual DGP), but is hard to avoid if the $\boldsymbol{\eta}_t$ are unknown. Conversely, when the GUM based on $\mathbf{w}_t$ nests the DGP, but the $\boldsymbol{\eta}_t$ are omitted from the theory model, or when selection with SIS detects DGP shifts outside the theory, then selection could greatly improve the final model, as in Castle and Hendry (2014a).

Consider a simple theory model which correctly matches its stationary DGP by specifying that:

$$y_t = \boldsymbol{\beta}'\mathbf{x}_t + \epsilon_t \tag{16}$$

where $\epsilon_t \sim \mathsf{IID}[0, \sigma_\epsilon^2]$ over $t = 1, \ldots, T$, and ϵ_t is independent of the n strongly exogenous variables $\{\mathbf{x}_1, \ldots, \mathbf{x}_t\}$, assumed to satisfy:

$$T^{-1}\sum_{t=1}^{T}\mathbf{x}_t\mathbf{x}_t' \xrightarrow{P} \boldsymbol{\Sigma}_{xx}$$

which is positive definite, so that:

$$T^{1/2}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\right) = \left(T^{-1}\sum_{t=1}^{T}\mathbf{x}_t\mathbf{x}_t'\right)^{-1} T^{-1/2}\sum_{t=1}^{T}\mathbf{x}_t\epsilon_t \xrightarrow{D} \mathsf{N}_n\left[\mathbf{0}, \sigma_\epsilon^2\boldsymbol{\Sigma}_{xx}^{-1}\right] \tag{17}$$

where $\boldsymbol{\beta}_0$ is the population value of the parameter.

To check her theory in (16) against more general specifications, an investigator postulates:

$$y_t = \boldsymbol{\beta}'\mathbf{x}_t + \boldsymbol{\gamma}'\mathbf{v}_t + \epsilon_t \tag{18}$$

which includes the additional k exogenous variables $\mathbf{v}_t$ although in fact $\boldsymbol{\gamma}_0 = \mathbf{0}$. The $\mathbf{v}_t$ can be variables known to be exogenous, functions of those, lagged variables in time series, and indicators for outliers or shifts, and we assume the same assumptions as above for $\{\epsilon_t, \mathbf{x}_t, \mathbf{v}_t\}$.[2] As she believes her theory in (16) is correct and complete, the $\mathbf{x}_t$ are always retained (so not selected over, called forced). We now consider the impact of selecting for significant candidate variables from $\mathbf{v}_t$ in (18), while retaining the $\mathbf{x}_t$, relative to directly estimating the DGP in (16) when $(n+k) \ll T$ to illustrate the approach.

First orthogonalize $\mathbf{v}_t$ and $\mathbf{x}_t$ by the projection:

$$\mathbf{v}_t = \Gamma\mathbf{x}_t + \mathbf{e}_t \tag{19}$$

so that:

$$\widehat{\boldsymbol{\Gamma}} = \left(\sum_{t=1}^{T} \mathbf{v}_t\mathbf{x}_t'\right)\left(\sum_{t=1}^{T} \mathbf{x}_t\mathbf{x}_t'\right)^{-1} \tag{20}$$

and letting:

$$\mathbf{e}_t = \mathbf{v}_t - \widehat{\boldsymbol{\Gamma}}\mathbf{x}_t \tag{21}$$

then by construction:

$$\sum_{t=1}^{T} \mathbf{e}_t\mathbf{x}_t' = \mathbf{0}. \tag{22}$$

Substituting $\mathbf{e}_t$ in (21) for $\mathbf{v}_t$ in (18) and using $\boldsymbol{\gamma}_0 = \mathbf{0}$, so that $\widehat{\boldsymbol{\Gamma}}'\boldsymbol{\gamma}_0 = \mathbf{0}$:

$$\begin{aligned} y_t &= \boldsymbol{\beta}'\mathbf{x}_t + \boldsymbol{\gamma}'\left(\widehat{\boldsymbol{\Gamma}}\mathbf{x}_t + \mathbf{e}_t\right) + \epsilon_t \\ &= (\boldsymbol{\beta} + \boldsymbol{\gamma}'\widehat{\boldsymbol{\Gamma}})'\mathbf{x}_t + \boldsymbol{\gamma}'\mathbf{e}_t + \epsilon_t = \boldsymbol{\beta}'\mathbf{x}_t + \boldsymbol{\gamma}'\mathbf{e}_t + \epsilon_t \end{aligned} \tag{23}$$

noting $\widehat{\boldsymbol{\Gamma}}'\boldsymbol{\gamma} = \widehat{\boldsymbol{\Gamma}}'\boldsymbol{\gamma}_0 = \mathbf{0}$, but retaining the $\mathbf{e}_t$ as regressors. As is well known in orthogonal regressions, from (23), when (16) is the DGP and using (22):

$$\begin{aligned} \sqrt{T}\begin{pmatrix}\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 \\ \widetilde{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0\end{pmatrix} &= \begin{pmatrix} T^{-1}\sum_{t=1}^{T}\mathbf{x}_t\mathbf{x}_t' & T^{-1}\sum_{t=1}^{T}\mathbf{x}_t\mathbf{e}_t' \\ T^{-1}\sum_{t=1}^{T}\mathbf{e}_t\mathbf{x}_t' & T^{-1}\sum_{t=1}^{T}\mathbf{e}_t\mathbf{e}_t'\end{pmatrix}^{-1} \\ &\times \begin{pmatrix} T^{-1/2}\sum_{t=1}^{T}\mathbf{x}_t\epsilon_t \\ T^{-1/2}\sum_{t=1}^{T}\mathbf{e}_t\epsilon_t\end{pmatrix} \\ &\xrightarrow{D} \mathsf{N}_{n+k}\left[\begin{pmatrix}\mathbf{0}\\ \mathbf{0}\end{pmatrix}, \sigma_\epsilon^2\begin{pmatrix}\boldsymbol{\Sigma}_{xx}^{-1} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_{vv|x}^{-1}\end{pmatrix}\right] \end{aligned} \tag{24}$$

where $\boldsymbol{\Sigma}_{vv|x}$ is the plim of the conditional variance of $\mathbf{v}_t$ given $\mathbf{x}_t$. From (24), the distribution of the estimator $\widetilde{\boldsymbol{\beta}}$ is identical to that of $\widehat{\boldsymbol{\beta}}$ in (17), irrespective of including or excluding any of the $\mathbf{e}_t$. Thus, whether or not selection of $\mathbf{e}_t$ is undertaken by a procedure that only retains estimated $\widetilde{\gamma}_i$ that are significant on some criterion (e.g., at significance level α, and corresponding critical value c_α, say, by sequential t-tests) will not affect the distribution of $\widetilde{\boldsymbol{\beta}}$. In fact that statement remains true even if $\boldsymbol{\gamma}_0 \neq \mathbf{0}$, although then both $\widetilde{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\beta}}$ are centered on $\boldsymbol{\beta}_0 + \boldsymbol{\Gamma}'\boldsymbol{\gamma}_0$, rather than $\boldsymbol{\beta}_0$.

This approach to model selection by merging theory-driven (i.e., always retain the theory model unaffected by selection) and data-driven (i.e., search over a large number of potentially relevant candidate variables) seems like a free lunch: the distribution of the theory parameters' estimates is the same with or without selection, so search is costless under the null. Moreover, the theory model is tested simultaneously against a wide range of alternatives represented by $\mathbf{v}_t$, enhancing its credibility if it is not rejected—in essence finding the answer to every likely seminar question in advance while controlling for false positives.

The main 'problem' is deciding when the theory model is rejected, since under the null, some of the $\widetilde{\gamma}_i$ will be significant by chance. When $\boldsymbol{\gamma}_0 = \mathbf{0}$, so all $e_{i,t}$ are irrelevant, nevertheless on average αk of the $e_{i,t}$ will be significant by chance at c_α using (say) selection by the rule:

$$\left|\mathrm{t}_{\gamma_i=0}\right| = \frac{|\widetilde{\gamma}_i|}{\mathsf{SE}\left[\widetilde{\gamma}_i\right]} \geq c_\alpha. \tag{25}$$

Hendry and Johansen (2015) propose setting $\alpha = \min(1/k, 1/T, 1\%)$ so, e.g., at $T = 500$ with $\alpha = 0.002$, even for $k = 100$, then $k\alpha = 0.2$, and the probability of retaining more than two irrelevant variables, r, is:

$$\begin{aligned} \Pr\{r > 2\} &= 1 - \sum_{i=0}^{2}\frac{(k\alpha)^i}{i!}e^{-k\alpha} = 1 - \sum_{i=0}^{2}\frac{(0.2)^i}{i!}e^{-0.2} \\ &\simeq 0.1\%. \end{aligned} \tag{26}$$

Hence, finding that 3 or more of the 100 $\widehat{e}_{i,t}$ are significant at 0.2% would strongly reject the null. Clearly, (26) can be evaluated for other choices of critical values, or a joint F-test could be used, but it seems reasonable to require relatively stringent criteria to reject the incumbent theory.

An unbiased estimator of the equation error variance under the null that $\boldsymbol{\gamma}_0 = \mathbf{0}$ is given by:

$$\widetilde{\sigma}_\epsilon^2 = (T-n)^{-1}\sum_{t=1}^{T}\left(y_t - \widetilde{\boldsymbol{\beta}}'\mathbf{x}_t\right)^2. \tag{27}$$

Estimates of selected γ_i can also be approximately bias corrected, as in Castle, Doornik, and Hendry (2016) and Hendry and Krolzig (2005), who show this reduces their RMSEs under the null of irrelevance.

When the theory is incomplete or incorrect, the model selected from this approach will be less mis-specified than direct estimation, so we now discuss this win-win outcome.

[2] Hendry and Johansen (2015) also consider instrumental variables variants when some of the regressors are endogenous, and note that the validity of over-identified instruments can be checked as in Durbin (1954), Hausman (1978), and Wu (1973).

7.4. Working with an incomplete or invalid theory

First we consider when (18) nests a DGP where some of the $\mathbf{v}_t$ are relevant and some of the $\mathbf{x}_t$ are irrelevant. Under the alternative that $\boldsymbol{\gamma}_0 \neq \mathbf{0}$, directly fitting (16) will result in estimating the population coefficient $\boldsymbol{\beta}_0 + \boldsymbol{\Gamma}'\boldsymbol{\gamma}_0$. That will also be the coefficient of $\mathbf{x}_t$ in (23) when (18) nests a more general DGP, so as before the distributions of $\widehat{\boldsymbol{\beta}}$ from (17) and $\widetilde{\boldsymbol{\beta}}$ from (24) are the same. However, now the relevant components of $\mathbf{e}_t$ are likely to be significant, not only leading to rejection of the theory model as incomplete or incorrect, but also delivering an improved model of the DGP that reveals which additional combinations of variables matter. Despite retaining the theory-based $\mathbf{x}_t$ without selection in (23), that does not ensure the significance of its coefficients when the DGP is more general than the theory model. That may not be obvious from simply fitting (16), so to understand why elements of $\mathbf{x}_t$ are insignificant in the enlarged specification and which $v_{i,t}$ matter will usually require selecting from (18).

When (18) does not nest the DGP, but some $\mathbf{v}_t$ are relevant, selection from (23) will still usually improve the final model relative to (16), as in Castle et al. (2011). Providing relevant $v_{i,t}$ are then retained, the theory model will be rejected in favour of a more general formulation. Retaining $\mathbf{x}_t$ without selection will again deliver an estimate of $\boldsymbol{\beta}_0$ confounded with $\boldsymbol{\Gamma}'\boldsymbol{\gamma}_0$, but as in the previous paragraph this can be untangled if required. In both cases, the inadequacy of (16) results.

Hendry and Johansen (2015) also show that the above results generalize to a setting where there are more candidate variables $n + k$ than observations T, based on the analysis in Johansen and Nielsen (2009), providing $n \ll T$. Orthogonalizing the $\mathbf{x}_t$ with respect to the other candidate variables needs to be done in feasibly sized blocks, and selection then undertaken using a combination of expanding and contracting multiple block searches as implemented in (e.g.) *Autometrics*: see Doornik (2009) and Doornik and Hendry (2013).

While this discussion addresses the simple case of linear regression, the approach to retaining a theory model embedded in a set of candidate orthogonalized variables is general, and could be extended to a wide range of settings, including systems. Under the null, there will be no impact on the estimated parameter distributions relative to direct estimation, but when the extended model nests the DGP, improved results will be obtained. If the initial theory model was believed to be 'best practice', then new knowledge will have been discovered as discussed by Hendry and Doornik (2014).

8. Evaluating policy models

Many empirical macroeconomic models are built to provide policy analyses. Usually such models emphasize theory-based specifications, even to the extent of retaining them despite being rejected against the available data. Hopefully the approach in the previous section will help provide a better balance between theory consistency and verisimilitude, as location shifts are a crucial feature absent from most theory-based models, yet can have detrimental impacts on policy analyses as we now discuss.

When location shifts occur, estimated models are constant only if:
(a) all substantively relevant variables are included; and
(b) none of the shifts are internal to the model,
both of which are demanding requirements. One can tackle (a) by commencing from a large initial information set as in Section 7 without necessarily impugning the theory basis when it is viable. For (b), appropriate indicators for location shifts would remove the non-constancy, as will IIS or SIS when they accurately match such indicators. Otherwise, omissions of substantively relevant variables entail that policy derivatives will be incorrect. Castle and Hendry (2014a) analyze models that unknowingly omit variables which are correlated with included regressors when location shifts occur in both sets. They show that shifts in omitted variables lead to location shifts in the model, but do not alter estimated slope parameters. Surprisingly, location shifts in the included variables change both model intercepts and alter estimated slope parameters. Thus, unless policy models have no omitted substantively-relevant variables correlated with included variables, or the policy change is just a mean-preserving spread, policy changes will alter the estimated model, so the outcome will differ from the anticipated effect. The policy model becomes non-constant precisely when the policy is changed, which is a highly detrimental failure of super exogeneity. Thus, a test for super exogeneity prior to policy implementation is essential.

It is possible to investigate in advance if any in-sample policy changes had shifted the marginal processes of a model, and check whether these coincide with shifts in the relevant conditional model, so implementing co-breaking. The test in Hendry and Santos (2010) is based on IIS in the marginal models, retaining all the significant outcomes and testing their relevance in the conditional model. No *ex ante* knowledge of timings or magnitudes of breaks is required, and one need not know the DGP of the marginal variables. Nevertheless, the test has the correct size under the null of super exogeneity for a range of sizes of marginal-model saturation tests. The test has power to detect failures of super exogeneity when there are location shifts in any marginal models, so applies as well to models with expectations, like new-Keynesian Phillips curves (NKPCs): see Castle et al. (2014).

A generalization of that test to SIS is described in Castle, Hendry, and Martinez (2017) as follows. The first stage applies SIS to the marginal system, retaining indicators at significance level α_1:

$$\mathbf{x}_t = \boldsymbol{\pi}_0 + \sum_{j=1}^{s} \boldsymbol{\Pi}_j \mathbf{x}_{t-j} + \sum_{i=1}^{m} \boldsymbol{\rho}_{i,\alpha_1} 1_{\{t \leq t_i\}} + \mathbf{e}_t. \tag{28}$$

The second stage adds the m retained step indicators to the conditional equation, say:

$$y_t = \mu_0 + \boldsymbol{\beta}'\mathbf{x}_t + \sum_{i=1}^{m} \tau_{i,\alpha_2} 1_{\{t \leq t_i\}} + \epsilon_t \tag{29}$$

and conducts an F-test for the significance of the $\left(\tau_{1,\alpha_2} \dots \tau_{m,\alpha_2}\right)$ at level α_2. The test has power as significant

step indicators capture outliers and location shifts that are not otherwise explained by the regressors, so super exogeneity does not then hold. Given a rejection, a rethink of the policy model seems advisable.

9. Nowcasts and flash estimates

Nowcasts, 'forecasts of the present state of the economy', usually based on high-frequency or mixed-data sampling, and flash estimates 'calculating what an *aggregate* outcome will eventually be based on a subset of disaggregate information', are both in common use for establishing the value of the forecast origin, critical for forecasting and policy. See among many others, Angelini, Camba-Méndez, Giannone, Rünstler, and Reichlin (2011), Doz, Giannone, and Reichlin (2011), Ghysels, Sinko, and Valkanov (2007), Giannone, Reichlin, and Small (2008), Stock and Watson (2002) and the survey in Bánbura, Giannone, and Reichlin (2011) for the former, and e.g., Castle, Fawcett, and Hendry (2009) and many of the chapters in EuroStat (in press) for the latter. Although based on different information sets and approaches, both are subject to many of the same difficulties that confront forecasting, as discussed in Castle, Hendry, and Kitov (in press).

'Information' is an ambiguous concept in a wide-sense non-stationary world, about which knowledge is limited, yet needed for rational action. In the theory of information originating in the seminal research of Shannon (1948), semantic aspects are irrelevant to processing message signals. But meaning matters in economics, so there are four different usages of 'information', and they play different roles. To illustrate, consider seeking to estimate a conditional expectation a short time ahead, given by $\mathsf{E}_{\mathsf{D}_{y_{t+1}}}[y_{t+1}|\mathcal{J}_{t+\delta}]$ (say) where $1 > \delta > 0$, using the 'information' set $\mathcal{J}_{t+\delta}$ when the DGP at $t+1$ depends on $\mathcal{I}_{t+1} \supseteq \mathcal{J}_{t+1}$:

[a] 'information' meaning knowing the variables that generate the σ-field $\mathcal{J}_{t+\delta}$, say a set $\{\mathbf{x}_{t+\delta}\}$;

[b] 'information' meaning knowledge of the form and structure of the distribution $\mathsf{D}_{y_{t+1}}(\cdot)$;

[c] 'information' meaning knowing how $\mathcal{J}_{t+\delta}$ enters the conditional distribution $\mathsf{D}_{y_{t+1}|\mathcal{J}_{t+\delta}}(y_{t+1}|\mathcal{J}_{t+\delta})$ (i.e., linearly or not; levels or differences; at what lags; needing what deterministic terms, etc.);

[d] 'information' meaning knowing how $\mathsf{D}_{y_{t+1}}(\cdot)$ changed in the next period from $\mathsf{D}_{y_t}(\cdot)$, essential even when the latter is 'known' (but probably just estimated);

[e] 'information' about the components of $\mathcal{I}_{t+\delta} - \mathcal{J}_{t+\delta}$.

In practice, although nowcasting concerns the present state of the economy, neither $\mathcal{I}_{t+1}$ nor $\mathcal{J}_{t+1}$ are usually known at time $t+1$, so y_{t+1} has to be inferred, rather than calculated from available disaggregate information, leading to the formulation $\mathsf{E}_{\mathsf{D}_{y_t}}[y_{t+1}|\mathcal{J}_{t+\delta}]$ as the 'best' estimate of y_{t+1}. Notice that only $\mathsf{D}_{y_t}(\cdot)$ will be available, but conversely, high-frequency information on some components of $\{\mathbf{x}\}$ may be available at $t+\delta$. Given this framing, and the five senses [a]–[e] just delineated, we can consider the roles the approaches in Sections 4–8 might play in nowcasting, focusing on the strategic aspects rather than the tactics of how to handle important but detailed issues such as missing data, measurement errors, 'ragged edges' etc.

First, [a] is often the sense of 'information', leading to 'big data' approaches based on principal components or estimated factors of very large cross sections of variables, so are primarily empirically based, although smaller, more focused information sets have also been used in the nowcasting context: see e.g., Camacho and Perez-Quiros (2010). When location shifts can occur, a potential drawback of simply pooling by averaging over huge data sets, as with say, principal components, is that a very large change in what was once a useful element can make it become poisonous. A recent example is the dramatic increase in the balance sheet of the Fed after quantitative easing: without selecting out such a variable, forecast, and nowcast, failure would occur. A potential advantage when the data involved are high frequency is early detection of shifts in $\mathsf{D}_{y_{t+1}}(\cdot)$, which could be found by testing for forecast failure at every period $t+\delta$, where (e.g.) δ denotes a monthly observation for forecasting a quarterly aggregate, attempting to resolve [d] as well. Conversely, economic theory based approaches essentially claim to provide solutions to [b] and [c], postulating the distribution, the subset of relevant variables as 'information', and how they enter $\mathsf{D}_{y_{t+1}}(\cdot)$. The recent track record of macroeconomic model-based forecasts suggests those postulates do not closely match reality, as with the Bank of England's COMPASS model noted above. The implications for policy, and for basing 'stories' about the future behaviour of economies purely on such models are all too obvious. In addition to the approach discussed in Section 8, policy models can be evaluated by checking the validity of the accompanying narratives often used to justify their forecasts and policies. Castle et al. (2017) call this process **forediction** to capture the notion of **fore**casts with stories (**dictions**), following the joint findings of Ericsson (2016) and Stekler and Symington (2016) that remarkably accurate reconstructions of Federal Open Market Committee (FOMC) forecasts can be derived from the verbal records in minutes of their meetings. Indeed, the more closely the policy story and forecasts are tied, the more likely is forecast failure to entail forediction failure, revealing the stories to be fiction, with possibly policy invalidity, noting the key caveat that temporary evidence of forecast failure due to poor forecast-origin values might be reversed in the next period, an issue nowcasting seeks to avoid. However, the failure of the law of iterated expectations from unanticipated distributional shifts makes multistep forecasting, as required by many policy agencies, even more hazardous than just 1-step ahead. Finally [e], if the missing 'information' would predict any shifts to $\mathsf{D}_{y_{t+1}}(\cdot)$, it would be invaluable. Conversely, if that 'information' merely concerned finding variables that entered the DGP with tiny constant effects, it would be essentially irrelevant, and might even improve nowcasts by being ignored, as with the trade-off in Akaike (1973) between the mis-specification costs of dropping small effects or the estimation costs of retaining them. Seeking to sort substantively relevant variables from others in the initial candidate set, as well as discovering what outliers and shifts occurred in-sample, and how variables enter the conditional distribution, is precisely where model selection enters, as an attempt to tackle all of [a]–[e]

jointly. Hopefully, some economic theories are sufficiently insightful to add value to that approach, as advocated in Section 7.

10. Conclusions

The theory of reduction provides an interpretation for the 'design of experiments' in Haavelmo (1944), whereby the local data generation process (LDGP) is intrinsically designed to match the 'passive observations of reality', so is the appropriate target for model selection. The relevant subject-matter theory remains the object of interest, and both specifies the initial set of relevant variables for the LDGP, and is embedded within the model of the LDGP. As the LDGP is always unknown in practice, it must be discovered from the available evidence. To do so, we seek to nest the LDGP so far as possible in a general unrestricted model (GUM), allowing for non-stationarities, especially location shifts, dynamic reactions, non-linearities and other potentially relevant variables, hence often leading to more candidate variables, N, than observations, T, while embedding the theory model within that GUM. Selection then seeks the simplest acceptable representation thereof, stringently evaluating it for congruence and encompassing

Against that background, alternative approaches that were purely data-driven or theory-driven were discussed. The economy is far too high-dimensional and inter-dependent to be modelled empirically without some prior insights into likely connections. However, the location shifts that perturb empirical models and lead to forecast failure also impact adversely on any 'theory only' approach. Location shifts invalidate the law of iterated expectations, and entail that conditional expectations based on the pre-existing distribution are not unbiased for the outcome in the next period. This not only invalidates the mathematical basis for inter-temporal derivations, the assumptions in such derivations could not correspond to economic agents' behaviour, unless agents are irrational. Following an unanticipated shift, agents could not quickly learn what it had changed, so would need *error-correction* mechanisms to avoid systematic mistakes. When policy models are not a complete specification of all relevant variables, policy changes can cause estimated slope parameters to alter and thereby induce a failure of super exogeneity.

Forecast failure is also primarily due to location shifts that were not included in the forecast, which unfortunately seems to be all too common. Consequently, all *equilibrium-correction* models then systematically mis-forecast, namely regressions, VARs, DSGEs, EqCMs, and GARCH representations. Conversely, every DGP parameter can shift without any noticeable effect on that class of model if there is no induced location shift. While it has proved very difficult to predict location shifts, it is possible to partially mitigate the systematic failure that follows by using robust forecasting devices. Thus, good (bad) forecast performance is not directly indicative of the verisimilitude (invalidity) of a forecasting device.

None of these three approaches is adequate in isolation for economies with high-dimensional, wide-sense non-stationary DGPs, requiring the estimation of models of evolving systems from relatively short, inter-dependent and heterogeneous aggregate time-series data, often subject to extensive revision, facing additional problems of endogeneity, potential lack of identification, and collinearity. However, location shifts are not purely detrimental: Duffy and Hendry (2017) show that they can help reveal substantive relationships between variables that are badly measured. Hence, given that the LDGP is the target for model selection once the basic formulation is enunciated, empirical model discovery with theory evaluation seems the only plausible alternative. It is straightforward to automatically create a very general model from an investigator's specification of the contents of $\mathbf{x}_t$, including additional variables, longer lags, non-linearities, and indicators for outliers and shifts to ensure congruent representations that might nest the LDGP, and avoid the potentially pernicious costs of under-specification. The theory-model can be embedded therein, orthogonalizing other candidate variables with respect to $\mathbf{x}_t$ to ensure that the baseline specification is retained unaltered. From this GUM, select the most parsimonious, congruent, encompassing model at the chosen significance level to ensure an undominated selection, controlling the probability of adventitiously significant findings at the desired level, computing bias-corrected parameter estimates for appropriate quantification. Stringently evaluate the results, especially super exogeneity, to ensure the selected model is robust to policy changes. The approach generalizes to $N > T$, with expanding and contracting searches. Surprisingly, empirical modelling is feasible in such a setting, even when pure theory models, forecasts and policy fail from location shifts. Empirical model discovery with theory evaluation offers a way ahead.

Acknowledgments

Financial support from the Robertson Foundation (grant 9907422), Institute for New Economic Thinking (grant 20029822) and Statistics Norway (through Research Council of Norway Grant 236935) is gratefully acknowledged, as are helpful comments from Jennifer L. Castle, Jurgen A. Doornik, Robert F. Engle, Neil R. Ericsson, Tito Nicias Teixeira da Silva Filho, Kevin Hoover, Katarina Juselius, Andrew B. Martinez, Grayham E. Mizon, John N.J. Muellbauer, Gabriel Perez-Quiros, and Tommaso Proietti, as well as participants in the ESRC–OMS 2012 International Symposium on Macroeconomics, the Ken@75 Conference, the Carlos III University Workshop Honoring Antoni Espasa, Department of Economics University of Oslo and the 2015 Haavelmo Memorial Lecture. The original impetus for the paper was the *Journal of Economic Surveys* Online Conference in November, 2011: see https://joesonlineconference.wordpress.com/. All calculations and graphs use *OxMetrics* (Doornik, 2013) and *PcGive* (Doornik and Hendry, 2013).

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov, & F. Csaki (Eds.), *Second international symposion on information theory* (pp. 267–281). Budapest: Akademia Kiado.

Angelini, E., Camba-Méndez, G., Giannone, D., Rünstler, G., & Reichlin, L. (2011). Short-term forecasts of Euro-area GDP growth. *The Econometrics Journal*, *14*, C25–C44.

Anundsen, A. K., Nymoen, R., Krogh, T. S., & Vislie, J. (2012). The macroeconomics of Trygve Haavelmo. *Nordic Journal of Political Economy*, *37*, 1–26.

Bánbura, M., Giannone, D., & Reichlin, L. (2011). Nowcasting. In M. P. Clements, & D. F. Hendry (Eds.), *Oxford handbook of economic forecasting*. Oxford: Oxford University Press, (Chapter 7).

Bjerkholt, O. (2005). Frisch's econometric laboratory and the rise of Trygve Haavelmo's probability approach. *Econometric Theory*, *21*, 491–533.

Bjerkholt, O. (2007). Writing 'the probability approach' with nowhere to go: Haavelmo in the United States, 1939-1944. *Econometric Theory*, *23*, 775–837.

Bontemps, C., & Mizon, G. E. (2008). Encompassing: Concepts and implementation. *Oxford Bulletin of Economics and Statistics*, *70*, 721–750.

Boumans, M. A. (1999). Representation and stability in testing and measuring rational expectations. *Journal of Economic Methodology*, *6*, 381–401.

Boumans, M. A., & Morgan, M. S. (2001). Ceteris paribus conditions: Materiality and the applications of economic theories. *Journal of Economic Methodology*, *8*, 11–26.

Burgess, S., Fernandez-Corugedo, E., Groth, C., Harrison, R., Monti, F., & Theodoridis, K. et al., (2013) The Bank of England's Forecasting Platform: COMPASS, MAPS, EASE and the suite of models. Working Paper 471 and Appendices, Bank of England, London.

Camacho, M., & Perez-Quiros, G. (2010). Introducing the EURO-STING: Short term indicator of euro area growth. *Journal of Applied Econometrics*, *25*, 663–694.

Castle, J. L., Clements, M. P., & Hendry, D. F. (2015). Robust approaches to forecasting. *International Journal of Forecasting*, *31*, 99–112.

Castle, J. L., Doornik, J. A., & Hendry, D. F. (2011). Evaluating automatic model selection. *Journal of Time Series Econometrics*, *3*(1), http://dx.doi.org/10.2202/1941-1928.1097.

Castle, J. L., Doornik, J. A., & Hendry, D. F. (2016). Automatic model selection with soft thresholding. Working Paper, Economics Department, Oxford University.

Castle, J. L., Doornik, J. A., Hendry, D. F., & Nymoen, R. (2014). Misspecification testing: Non-invariance of expectations models of inflation. *Econometric Reviews*, *33*, 553–574.

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

Castle, J. L., Fawcett, N. W. P., & Hendry, D. F. (2009). Nowcasting is not just contemporaneous forecasting. *National Institute Economic Review*, *210*, 71–89.

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

Castle, J. L., & Hendry, D. F. (2014a). Model selection in under-specified equations with breaks. *Journal of Econometrics*, *178*, 286–293.

Castle, J. L., & Hendry, D. F. (2014b). Semi-automatic non-linear model selection. In N. Haldrup, M. Meitz, & P. Saikkonen (Eds.), *Essays in nonlinear time series econometrics* (pp. 163–197). Oxford: Oxford University Press.

Castle, J. L., Hendry, D. F., & Kitov, O. I. (2014). Forecasting and nowcasting macroeconomic variables: A methodological overview. See EuroStat (in press).

Castle, J. L., Hendry, D. F., & Martinez, A. B. (2017). Evaluating forecasts, narratives and policy using a test of invariance. *Econometrics*, *5*(39), http://dx.doi.org/10.3390/econometrics5030039.

Castle, J. L., Qin, X., & Reed, W. R. (2013). Using model selection algorithms to obtain reliable coefficient estimates. *Journal of Economic Surveys*, *27*, 269–296.

Castle, J. L., & Shephard, N. (Eds.). (2009). *The methodology and practice of econometrics*. Oxford: Oxford University Press.

Clements, M. P., & Hendry, D. F. (1998). *Forecasting economic time series*. Cambridge: Cambridge University Press.

Clements, M. P., & Hendry, D. F. (1999). *Forecasting non-stationary economic time series*. Cambridge, Mass: MIT Press.

Doob, J. L. (1953). *Stochastic processes*. (1990th ed.). New York: John Wiley Classics Library.

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

Doornik, J. A. (2013). *OxMetrics: An interface to empirical modelling*. (7th ed.). London: Timberlake Consultants Press.

Doornik, J. A., & Hendry, D. F. (2013). *Empirical econometric modelling using PcGive: Vol. I*. (7th ed.). London: Timberlake Consultants Press.

Doornik, J. A., & Hendry, D. F. (2015). Statistical model selection with big data. *Cogent Economics and Finance*. http://dx.doi.org/10.1080/23322039.2015.1045216.

Doz, C., Giannone, D., & Reichlin, L. (2011). A two-step estimator for large approximate dynamic factor models based on Kalman filtering. *Journal of Econometrics*, *164*, 188–205.

Duffy, J. A., & Hendry, D. F. (2017). The impact of near-integrated measurement errors on modelling long-run macroeconomic time series. *Econometric Reviews*, *36*, 568–587.

Durbin, J. (1954). Errors in variables. *Review of the Institute of International Statistics*, *22*, 23–54.

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

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

EuroStat, (Eds.), (2017). Principal European economic indicators: Handbook on rapid estimates, Brussels, UN/EuroStat. (in press).

Ghysels, E., Sinko, A., & Valkanov, R. (2007). MIDAS regressions: Further results and new directions. *Econometric Reviews*, *26*, 53–90.

Giannone, D., Reichlin, L., & Small, D. (2008). Nowcasting GDP and inflation: The real-time informational content of macroeconomic data. *Journal of Monetary Economics*, *55*, 665–676.

Granger, C. W. J., & Teräsvirta, T. (1993). *Modelling nonlinear economic relationships*. Oxford: Oxford University Press.

Haavelmo, T. (1944). The probability approach in econometrics. *Econometrica*, *12*, 1–118 Supplement.

Haavelmo, T. (1958). The role of the econometrician in the advancement of economic theory. *Econometrica*, *26*, 351–357.

Haavelmo, T. (1989). Prize lecture. Sveriges Riksbank: Alfred Nobel Memorial Prize.

Harrison, R., Nikolov, M., Quinn, G., Ramsay, A., Scott, K., & Thomas, R. (2005). The Bank of England quarterly model. Research paper, pp. 244, Bank of England, London.

Hausman, J. A. (1978). Specification tests in econometrics. *Econometrica*, *46*, 1251–1271.

Hendry, D. F. (1987). Econometric methodology: A personal perspective. In T. F Bewley (Ed.), *Advances in econometrics* (pp. 29–48). Cambridge: Cambridge University Press.

Hendry, D. F. (2009). The methodology of empirical econometric modeling: Applied econometrics through the looking-glass. In T. C. Mills, & K. D. Patterson (Eds.), *Palgrave handbook of econometrics* (pp. 3–67). Basingstoke: Palgrave MacMillan.

Hendry, D. F. (2015). Introductory macro-econometrics: A new approach. London: Timberlake consultants. http://www.timberlake.co.uk/macroeconometrics.html.

Hendry, D. F., & Doornik, J. A. (2014). *Empirical model discovery and theory evaluation*. Cambridge, Mass: MIT Press.

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

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

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

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

Hendry, D. F., & Mizon, G. E. (2012). Open-model forecast-error taxonomies. In X. Chen, & N. R. Swanson (Eds.), *Recent advances and future directions in causality, prediction, and specification analysis* (pp. 219–240). New York: Springer.

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

Hendry, D. F., & Morgan, M. S. (Eds.). (1995). *The foundations of econometric analysis*. Cambridge: Cambridge University Press.

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

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

Johansen, S., & Nielsen, B. (2016). Asymptotic theory of outlier detection algorithms for linear time series regression models. *Scandinavian Journal of Statistics*, *43*, 321–348.
Juselius, K. (1993). VAR modelling and Haavelmo's probability approach to econometrics. *Empirical Economics*, *18*, 595–622.
Leeb, H., & Pötscher, B. M. (2003). The finite-sample distribution of post-model-selection estimators, and uniform versus non-uniform approximations. *Econometric Theory*, *19*, 100–142.
Leeb, H., & Pötscher, B. M. (2005). Model selection and inference: Facts and fiction. *Econometric Theory*, *21*, 21–59.
Mizon, G. E. (1995). A simple message for autocorrelation correctors: Don't. *Journal of Econometrics*, *69*, 267–288.
Mizon, G. E., & Richard, J.-F. (1986). The encompassing principle and its application to non-nested hypothesis tests. *Econometrica*, *54*, 657–678.
Morgan, M. S. (1990). *The history of econometric ideas*. Cambridge: Cambridge University Press.
Pretis, F., Reade, J. J., & Sucarrat, G. (2017). Automated general-to-specific (GETS) modelling of the mean and variance of regressions, and indicator saturation methods for outliers and structural breaks. *Journal of Statistical Software* (in press).
Qin, D. (1993). *The formation of econometrics: A historical perspective*. Oxford: Clarendon Press.
Qin, D. (2013). *A history of econometrics: The reformation from the 1970s*. Oxford: Clarendon Press.
Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, *27*, 623–656.
Smith, A. (1759). *Theory of moral sentiments*. Edinburgh: A. Kincaid, J. Bell.
Smith, B. B. (1929). Judging the forecast for 1929. *Journal of the American Statistical Association*, *24*, 94–98.
Soros, G. (2008). *The new paradigm for financial markets*. London: Perseus Books.
Spanos, A. (1989). On re-reading Haavelmo: A retrospective view of econometric modeling. *Econometric Theory*, *5*, 405–429.
Stekler, H. O., & Symington, H. (2016). Evaluating qualitative forecasts: The FOMC minutes, 2006-2010. *International Journal of Forecasting*, *32*, 559–570.
Stock, J. H., & Watson, M. W. (1999). A comparison of linear and nonlinear models for forecasting macroeconomic time series. In R. F. Engle, & H. White (Eds.), *Cointegration, causality and forecasting: A festschrift in honour of Clive W.J. Granger* (pp. 1–44). Oxford: Oxford University Press.
Stock, J. H., & Watson, M. W. (2002). Macroeconomic forecasting using diffusion indices. *Journal of Business & Economic Statistics*, *20*, 147–162.
Taleb, N. N. (2007). *The black swan*. New York: Random House.
Wu, D. (1973). Alternative tests of independence between stochastic regressors and disturbances. *Econometrica*, *41*, 733–750.