Unpredictability and the Foundations of Economic Forecasting

David F. Hendry*
Department of Economics, Oxford University.

July 11, 2005

Abstract

We revisit the concept of unpredictability to explore its implications for forecasting strategies in a non-stationary world subject to structural breaks, where model and mechanism differ. Six aspects of the role of unpredictability are distinguished, compounding the four additional mistakes most likely in estimated forecasting models. Structural breaks, rather than limited information, are the key problem, exacerbated by conflicting requirements on 'forecast-error corrections'. We consider model transformations and corrections to reduce forecast-error biases, as usual at some cost in increased forecast-error variances. The analysis is illustrated by an empirical application to M1 in the UK.

Contents

*Preliminary and incomplete, prepared for the first ESF-EMM conference: please do not cite without the author's permission. Financial support from the ESRC under grant RES051270035, and helpful comments from Mike Clements, Neil Ericsson, and Grayham Mizon, are all gratefully acknowledged.

1 Introduction

The historical track record of econometric systems is both littered with forecast failures, and their empirical out-performance by 'naive devices': see, for example, many of the papers reprinted in Mills (1999). At first sight, such an adverse outcome for econometric systems is surprising: since they incorporate inter-temporal causal information representing inertial dynamics in the economy, such models should have smaller prediction errors than purely extrapolative devices—but do not. In fact, discussions of the problems confronting economic forecasting date from the early history of econometrics: see, *inter alia*, Persons (1924), Morgenstern (1928) and Marget (1929). To explain such outcomes, Clements and Hendry (1998, 1999) developed a theory of forecasting for non-stationary processes subject to structural breaks, where the forecasting model differed from the data generating mechanism (extended from a theory implicitly based on the assumptions that the model coincided with a constant-parameter mechanism). They thereby accounted for the successes and failures of various alternative forecasting approaches, and helped explain the outcomes of forecasting competitions (see e.g., Makridakis and Hibon, 2000, Clements and Hendry, 2001a, and Fildes and Ord, 2002).

Following Clements and Hendry (1996), consider T observations $\mathbf{X}_{\mathbf{1}}^{\mathbf{T}} = (\mathbf{x}_{\mathbf{1}}, \ldots, \mathbf{x}_{\mathbf{T}})$ on a vector random variable, from which to predict the H future values $\mathbf{X}_{\mathbf{T+1}}^{\mathbf{T+H}} = (\mathbf{x}_{\mathbf{T+1}}, \ldots, \mathbf{x}_{\mathbf{T+H}})$. The joint probability of the observed and future $\mathbf{x}$s is $\mathsf{D}_{\mathbf{X}_{\mathbf{1}}^{\mathbf{T+H}}}(\mathbf{X}_{\mathbf{1}}^{\mathbf{T+H}} | \mathbf{X}_{\mathbf{0}}, \boldsymbol{\theta})$ where $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subseteq \mathbb{R}^{\mathbf{p}}$ is the parameter vector, and $\mathbf{X}_{\mathbf{0}}$ denotes the initial conditions. Factorizing into conditional and marginal probabilities:

$$\mathsf{D}_{\mathbf{X}_{\mathbf{1}}^{\mathbf{T+H}}}\left(\mathbf{X}_{\mathbf{1}}^{\mathbf{T+H}} \mid \mathbf{X}_{\mathbf{0}}, \boldsymbol{\theta}\right) = \mathsf{D}_{\mathbf{X}_{\mathbf{T+1}}^{\mathbf{T+H}}}\left(\mathbf{X}_{\mathbf{T+1}}^{\mathbf{T+H}} \mid \mathbf{X}_{\mathbf{1}}^{\mathbf{T}}, \mathbf{X}_{\mathbf{0}}, \boldsymbol{\theta}\right) \times \mathsf{D}_{\mathbf{X}_{\mathbf{1}}^{\mathbf{T}}}\left(\mathbf{X}_{\mathbf{1}}^{\mathbf{T}} \mid \mathbf{X}_{\mathbf{0}}, \boldsymbol{\theta}\right). \tag{1}$$

$\mathsf{D}_{\mathbf{X}_{\mathbf{T+1}}^{\mathbf{T+H}}}(\cdot)$ is unknown, so must be derived from $\mathsf{D}_{\mathbf{X}_{\mathbf{1}}^{\mathbf{T}}}(\cdot)$, which requires the 'basic assumption' that:

> 'The probability law $\mathsf{D}_{\mathbf{X}_{\mathbf{1}}^{\mathbf{T+H}}}(\cdot)$ of the $T + H$ variables $(\mathbf{x}_{\mathbf{1}}, \ldots, \mathbf{x}_{\mathbf{T+H}})$ is of such a type that the specification of $\mathsf{D}_{\mathbf{X}_{\mathbf{1}}^{\mathbf{T}}}(\cdot)$ implies the complete specification of $\mathsf{D}_{\mathbf{X}_{\mathbf{1}}^{\mathbf{T+H}}}(\cdot)$ and, therefore, of $\mathsf{D}_{\mathbf{X}_{\mathbf{T+1}}^{\mathbf{T+H}}}(\cdot)$.' (Haavelmo, 1944, p.107: my notation).

This formulation highlights the major problems that need to be confronted for successful forecasting. The form of $\mathsf{D}_{\mathbf{X}_{\mathbf{1}}^{\mathbf{T}}}(\cdot)$ and the value of $\boldsymbol{\theta}$ in sample must be learned from the observed data, involving problems of: *specification* of the set of relevant variables $\{\mathbf{x}_{\mathbf{t}}\}$, *measurement* of the $\mathbf{x}$s, *formulation* of the joint density $\mathsf{D}_{\mathbf{X}_{\mathbf{1}}^{\mathbf{T}}}(\cdot)$, *modelling* of the relationships, and *estimation* of $\boldsymbol{\theta}$, all of which introduce uncertainties, the baseline level of which is set by the *properties* of $\mathsf{D}_{\mathbf{X}_{\mathbf{1}}^{\mathbf{T}}}(\cdot)$. When forecasting, $\mathsf{D}_{\mathbf{X}_{\mathbf{T+1}}^{\mathbf{T+H}}}(\cdot)$ determines the 'intrinsic' uncertainty, rapidly *growing* as H increases–especially for *non-stationary* data (from stochastic trends etc.)–further increased by any *changes* in the distribution function $\mathsf{D}_{\mathbf{X}_{\mathbf{T+1}}^{\mathbf{T+H}}}(\cdot)$ or parameters thereof between T and later (lack of time invariance). These ten italicised issues structured their analysis of economic forecasting, but they emphasised the importance of the last of these.

The complementary, 'bottom up' explanation proposed here lies in the many steps between the ability to predict a random variable at a point in time, and a forecast of the realizations of that variable over a future horizon from a model based on an historical sample. This paper spells out those steps,

and demonstrates that many of the results on forecasting in Clements and Hendry (1998, 1999) have a foundation in the properties of unpredictability.

Having established foundations for their findings in the concept of unpredictability, this paper draws some implications for forecasting non-stationary processes using incomplete (i.e., mis-specified) models. The objective of this analysis is to ascertain ways of implementing the strengths of so-called 'naive' methods in macro-econometric models, via a 'forecasting strategy' which uses a combination of their 'causal' information with a more 'robust' forecasting device. Such a combination could be either by rendering the econometric system robust, or by modifying a robust device using an estimate of any likely causal changes. This paper concerns the former: for the latter, in the policy context, see Hendry and Mizon (2000, 2003). Although combining forecasts has a long pedigree (see, e.g., Bates and Granger, 1969, Diebold and Pauly, 1987, Clemen, 1989, Diebold and Lopez, 1996, Stock and Watson, 1999, and Newbold and Harvey, 2002) and a theory for its success (see Granger, 1989, and Hendry and Clements, 2004), we consider instead transformations of econometric systems that may improve their performance in the face of structural breaks.

We first review the well-established concept of unpredictability in section 2 and the transformations under which it is invariant (based on Hendry, 1997), with extensions of earlier results to non-stationary processes. Then section 3 draws its implications for the formulation of forecasting devices. Section 4 specifies a cointegrated DGP subject to breaks, and section 5 examines some adaptive devices which might improve its robustness in forecasting. Section 6 illustrates the ideas for the much-used empirical example of the behaviour of UK M1. Finally, section 7 concludes.

2 Unpredictability: A review and extension

A non-degenerate vector random variable $\boldsymbol{\nu}_{\mathbf{t}}$ is an unpredictable process with respect to an information set $\mathcal{I}_{\sqcup-\infty}$ over a period $\mathcal{T}$ if its conditional distribution $\mathsf{D}_{\boldsymbol{\nu}_{\mathbf{t}}}(\boldsymbol{\nu}_{\mathbf{t}}|\mathcal{I}_{\sqcup-\infty})$ equals its unconditional $\mathsf{D}_{\boldsymbol{\nu}_{\mathbf{t}}}(\boldsymbol{\nu}_{\mathbf{t}})$:

$$\mathsf{D}_{\boldsymbol{\nu}_{\mathbf{t}}}(\boldsymbol{\nu}_{\mathbf{t}} \mid \mathcal{I}_{\sqcup-\infty}) = \mathsf{D}_{\boldsymbol{\nu}_{\mathbf{t}}}(\boldsymbol{\nu}_{\mathbf{t}}) \quad \forall \mathsf{t} \in \mathcal{T}. \tag{2}$$

Importantly, unpredictability is a property of $\boldsymbol{\nu}_{\mathbf{t}}$ in relation to $\mathcal{I}_{\sqcup-\infty}$ intrinsic to $\boldsymbol{\nu}_{\mathbf{t}}$, and not dependent on any aspect of our knowledge thereof: this is one of the key gaps between predictability, when (2) is false, to 'forecastability'. Note that $\mathcal{T}$ may be a singleton (i.e., $\{t\}$), and that $\mathcal{I}_{\sqcup-\infty}$ always includes the sigma-field generated by the past of $\boldsymbol{\nu}_{\mathbf{t}}$.

A necessary condition for (2) is that ν_t is unpredictable in mean (denoted E_{t}) and variance (denoted V_{t}) at each point in $\mathcal{T}$, so assuming the relevant moments exist:

$$\mathsf{E}_{\mathsf{t}}[\boldsymbol{\nu}_{\mathbf{t}} \mid \mathcal{I}_{\sqcup-\infty}] = \mathsf{E}_{\mathsf{t}}[\boldsymbol{\nu}_{\mathbf{t}}] \quad \text{and} \quad \mathsf{V}_{\mathsf{t}}[\boldsymbol{\nu}_{\mathbf{t}} \mid \mathcal{I}_{\sqcup-\infty}] = \mathsf{V}_{\mathsf{t}}[\boldsymbol{\nu}_{\mathbf{t}}]. \tag{3}$$

The former does not imply the latter (a predictable conditional mean with a randomly heteroscedastic variance), or vice versa (e.g., an autoregressive conditional heteroscedastic–ARCH–process, as in (7) below, affecting a martingale difference sequence). Throughout, we will take the mean of the unpredictable process to be zero: $\mathsf{E}_{\mathsf{t}}[\boldsymbol{\nu}_{\mathbf{t}}] = \mathbf{0}\ \forall t$. Since we will be concerned with the predictability of functions of $\boldsymbol{\nu}_{\mathbf{t}}$ and $\mathcal{I}_{\sqcup-\infty}$, such as (6) below, any mean otherwise present could be absorbed in the latter. Due to possible shifts in the underlying distributions, both the information set available and all expectations operators must be time dated, which anyway clarifies multi-step prediction as in $\mathsf{E}_{\mathsf{T+h}}[\boldsymbol{\nu}_{\mathbf{T+h}}|\mathcal{I}_{\mathcal{T}}]$ for $h > 1$. The paper will focus on the first two moments in (3), rather than the complete density in (2), although extensions to the latter are feasible (see e.g., Tay and Wallis, 2000): however, for normal distributions, (3) suffices.

Unpredictability is only invariant under non-singular contemporaneous transforms: inter-temporal transforms must affect predictability (so no unique measure of forecast accuracy exists: see e.g., Leitch and Tanner, 1991, Clements and Hendry, 1993, and Granger and Pesaran, 2000a, 2000b). Predictability therefore requires combinations with $\mathcal{I}_{\sqcup-\infty}$, as for example:

$$\mathbf{y_t} = \boldsymbol{\phi_t}\left(\mathcal{I}_{\sqcup-\infty}, \boldsymbol{\nu_t}\right) \tag{4}$$

so $\mathbf{y_t}$ depends on both the information set and the innovation component. Then:

$$\mathsf{D}_{\mathbf{y_t}}\left(\mathbf{y_t} \mid \mathcal{I}_{\sqcup-\infty}\right) \neq \mathsf{D}_{\mathbf{y_t}}\left(\mathbf{y_t}\right) \quad \forall \mathsf{t} \in \mathcal{T}. \tag{5}$$

Two special cases of (4) are probably the most relevant empirically in economics, namely (after appropriate data transformations, such as logs):

$$\mathbf{y_t} = \mathbf{f_t}\left(\mathcal{I}_{\sqcup-\infty}\right) + \boldsymbol{\nu_t} \tag{6}$$

and:

$$\mathbf{y_t} = \boldsymbol{\nu_t} \odot \boldsymbol{\varphi_t}\left(\mathcal{I}_{\sqcup-\infty}\right) \tag{7}$$

where $\odot$ denotes element by element multiplication, so that $y_{i,t} = \nu_{i,t}\varphi_{i,t}\left(\mathcal{I}_{\sqcup-\infty}\right)$. Combinations and generalizations of these are clearly feasible and are also potentially relevant.

In (6), $\mathbf{y_t}$ is predictable in mean even if $\boldsymbol{\nu_t}$ is not as:

$$\mathsf{E}_{\mathsf{t}}\left[\mathbf{y_t} \mid \mathcal{I}_{\sqcup-\infty}\right] = \mathbf{f_t}\left(\mathcal{I}_{\sqcup-\infty}\right) \neq \mathsf{E}_{\mathsf{t}}\left[\mathbf{y_t}\right],$$

in general. Thus, the 'events' which will help predict $\mathbf{y_t}$ in (6) must already have happened, and a forecaster 'merely' needs to ascertain what $\mathbf{f_t}\left(\mathcal{I}_{\sqcup-\infty}\right)$ comprises. The dependence of $\mathbf{y_t}$ on $\mathcal{I}_{\sqcup-\infty}$ could be indirect (e.g., own lags may 'capture' actual past causes) since systematic correlations over the relevant horizon could suffice for forecasting – if not for policy. However, such stable correlations are unlikely in economic time series (a point made by Koopmans, 1937). The converse to (6) in linear models is well known in terms of the prediction decomposition (sequential factorization) of the likelihood (see e.g., Schweppe, 1965): if a random variable $\mathbf{y_t}$ is predictable from $\mathcal{I}_{\sqcup-\infty}$, as in (6), then it can be decomposed into two orthogonal components, one of which is unpredictable on $\mathcal{I}_{\sqcup-\infty}$ (i.e., $\boldsymbol{\nu_t}$ here), so is a mean innovation. Since:

$$\mathsf{V}_{\mathsf{t}}\left[\mathbf{y_t} \mid \mathcal{I}_{\sqcup-\infty}\right] < \mathsf{V}_{\mathsf{t}}\left[\mathbf{y_t}\right] \quad \text{when} \quad \mathbf{f_t}\left(\mathcal{I}_{\sqcup-\infty}\right) \neq \mathbf{0} \tag{8}$$

predictability ensures a variance reduction, consistent with its nomenclature, since unpredictability entails equality from (8)—the 'smaller' the conditional variance matrix, the less uncertain is the prediction of $\mathbf{y_t}$ from $\mathcal{I}_{\sqcup-\infty}$.

Although $\mathbf{y_t}$ remains unpredictable in mean in (7):

$$\mathsf{E}_{\mathsf{t}}\left[\mathbf{y_t} \mid \mathcal{I}_{\sqcup-\infty}\right] = \mathsf{E}_{\mathsf{t}}\left[\boldsymbol{\nu_t} \odot \boldsymbol{\varphi_t}\left(\mathcal{I}_{\sqcup-\infty}\right) \mid \mathcal{I}_{\sqcup-\infty}\right] = \mathbf{0},$$

it is predictable in variance because:

$$\mathsf{E}_{\mathsf{t}}\left[\mathbf{y_t}\mathbf{y_t}' \mid \mathcal{I}_{\sqcup-\infty}\right] = \mathsf{E}_{\mathsf{t}}\left[\boldsymbol{\nu_t}\boldsymbol{\nu_t}' \odot \boldsymbol{\varphi_t}\left(\mathcal{I}_{\sqcup-\infty}\right)\boldsymbol{\varphi_t}\left(\mathcal{I}_{\sqcup-\infty}\right)' \mid \mathcal{I}_{\sqcup-\infty}\right] = \boldsymbol{\Omega}_{\boldsymbol{\nu_t}} \odot \boldsymbol{\varphi_t}\left(\mathcal{I}_{\sqcup-\infty}\right)\boldsymbol{\varphi_t}\left(\mathcal{I}_{\sqcup-\infty}\right)'.$$

A well known special case of (7) of considerable relevance in financial markets is when $\mathcal{I}_{\sqcup-\infty}$ is the sigma-field generated by the past of $\mathbf{y_t}$. For a scalar y_t with constant σ_v^2 and $\varphi\left(\cdot\right) = \sigma_t$, this yields:

$$y_t = \nu_t\sigma_t,$$

so that (G)ARCH processes are generated by (see e.g., Engle, 1982, and Bollerslev, 1986: Shephard, 1996, provides an excellent overview):

$$\sigma_t^2 = \varphi_0 + \sum_{i=1}^{p} \varphi_i y_{t-i}^2 + \sum_{j=1}^{p} \varphi_{p+j} \sigma_{t-j}. \tag{9}$$

Alternatively, $\varphi(\cdot) = \exp(\sigma_t/2)$ leads to stochastic volatility (here as a first-order process: see e.g., Taylor, 1986, Kim, Shephard and Chib, 1998 and again, Shephard, 1996):

$$\sigma_{t+1} = \varphi_0 + \varphi_1 \sigma_t + \eta_t. \tag{10}$$

In both classes of model (9) and (10), predictability of the variance can be important in its own right (e.g., pricing options as in Melino and Turnbull, 1990), or for deriving appropriate forecast intervals.

2.1 Prediction from a reduced information set

Predictability is obviously relative to the information set used—when $\mathcal{J}_{\sqcup -\infty} \subset \mathcal{I}_{\sqcup -\infty}$ it is possible that:

$$\mathsf{D}_{\mathbf{u_t}}(\mathbf{u_t} \mid \mathcal{J}_{\sqcup -\infty}) = \mathsf{D}_{\mathbf{u_t}}(\mathbf{u_t}) \text{ yet } \mathsf{D}_{\mathbf{u_t}}(\mathbf{u_t} \mid \mathcal{I}_{\sqcup -\infty}) \neq \mathsf{D}_{\mathbf{u_t}}(\mathbf{u_t}). \tag{11}$$

This result helps underpin both general-to-specific model selection and the related use of congruence as a basis for econometric modelling (see e.g., Hendry, 1995, and Bontemps and Mizon, 2003). In terms of the former, less is learned based on $\mathcal{J}_{\sqcup -\infty}$ than $\mathcal{I}_{\sqcup -\infty}$, and the variance (where it exists) of the unpredictable component is unnecessarily large. In terms of the latter, a later investigator may discover additional information in $\mathcal{I}_{\sqcup -\infty}$ beyond $\mathcal{J}_{\sqcup -\infty}$ which explains part of a previously unpredictable error.

Given the information set, $\mathcal{J}_{\sqcup -\infty} \subset \mathcal{I}_{\sqcup -\infty}$ when the process to be predicted is $\mathbf{y_t} = \mathbf{f_t}(\mathcal{I}_{\sqcup -\infty}) + \boldsymbol{\nu_t}$ as in (6), less accurate predictions will result, but they will remain unbiased. Since $\mathsf{E}_t[\boldsymbol{\nu_t} | \mathcal{I}_{\sqcup -\infty}] = \mathbf{0}$:

$$\mathsf{E}_t[\boldsymbol{\nu_t} \mid \mathcal{J}_{\sqcup -\infty}] = \mathbf{0},$$

so that:

$$\mathsf{E}_t[\mathbf{y_t} \mid \mathcal{J}_{\sqcup -\infty}] = \mathsf{E}_t[\mathbf{f_t}(\mathcal{I}_{\sqcup -\infty}) \mid \mathcal{J}_{\sqcup -\infty}] = \mathbf{g_t}(\mathcal{J}_{\sqcup -\infty}),$$

say. Let $\mathbf{e_t} = \mathbf{y_t} - \mathbf{g_t}(\mathcal{J}_{\sqcup -\infty})$, then, providing $\mathcal{J}_{\sqcup -\infty}$ is a proper information set containing the history of the process:

$$\mathsf{E}_t[\mathbf{e_t} \mid \mathcal{J}_{\sqcup -\infty}] = \mathbf{0},$$

so $\mathbf{e_t}$ is a mean innovation with respect to $\mathcal{J}_{\sqcup -\infty}$. However, as $\mathbf{e_t} = \boldsymbol{\nu_t} + \mathbf{f_t}(\mathcal{I}_{\sqcup -\infty}) - \mathbf{g_t}(\mathcal{J}_{\sqcup -\infty})$:

$$\mathsf{E}_t[\mathbf{e_t} \mid \mathcal{I}_{\sqcup -\infty}] = \mathbf{f_t}(\mathcal{I}_{\sqcup -\infty}) - \mathsf{E}_t[\mathbf{g_t}(\mathcal{J}_{\sqcup -\infty}) \mid \mathcal{I}_{\sqcup -\infty}] = \mathbf{f_t}(\mathcal{I}_{\sqcup -\infty}) - \mathbf{g_t}(\mathcal{J}_{\sqcup -\infty}) \neq \mathbf{0}.$$

As a consequence of this failure of $\mathbf{e_t}$ to be an innovation with respect to $\mathcal{I}_{\sqcup -\infty}$:

$$\mathsf{V}_t[\mathbf{e_t}] > \mathsf{V}_t[\boldsymbol{\nu_t}],$$

so less accurate predictions will result. Nevertheless, that predictions remain unbiased on the reduced information set suggests that, by itself, incomplete information is not fatal to the forecasting enterprise.

2.1.1 Changes in information sets

Similarly, predictability cannot increase as the horizon grows for a fixed event $\mathbf{y_T}$ based on $\mathcal{I}_{\mathcal{T}-\langle}$ for $h = 1, 2, \ldots, H$, since the information sets form a decreasing nested sequence going back in time:

$$\mathcal{I}_{\mathcal{T}-\mathcal{H}} \subseteq \mathcal{I}_{\mathcal{T}-\mathcal{H}+\infty} \subseteq \cdots \subseteq \mathcal{I}_{\mathcal{T}-\infty}. \tag{12}$$

Conversely, disaggregating components of $\mathcal{I}_{\mathcal{T}-\langle}$ into their elements cannot lower predictability of a given aggregate $\mathbf{y_T}$, where such disaggregation may be across space (e.g., regions of an economy), variables (such as sub-indices of a price measure), or both. Further, since a lower frequency is a subset of a higher, and unpredictability is not in general invariant to the data frequency, then (11) ensures that temporal disaggregation cannot lower the predictability of the same entity $\mathbf{y_T}$ (data frequency issues will reappear in section 3).

These attributes sustain general models, and so may provide a formal basis for including as much information as possible, being potentially consistent with many-variable 'factor forecasting' (see e.g. Stock and Watson, 1999, and Forni, Hallin, Lippi and Reichlin, 2000), and with the benefits claimed in the 'pooling of forecasts' literature (e.g., Clemen, 1989, and Hendry and Clements, 2004, for a recent theory). Although such results run strongly counter to the common finding in forecasting competitions that 'simple models do best' (see e.g., Makridakis and Hibon, 2000, Allen and Fildes, 2001, and Fildes and Ord, 2002), Clements and Hendry (2001a) suggest that simplicity is confounded with robustness, and there remains a large gap between predictability and forecasting, an issue addressed below.

In all these case, $\mathsf{D}_{\mathbf{y_{T+h}}}(\mathbf{y_{T+h}}|\cdot)$ remains the target of interest, and $\mathcal{I}_{\mathcal{T}-\langle}$ is 'decomposed', in that additional content is added to the information set. A different, but related, form of disaggregation is of the target variable $\mathbf{y_T}$ into its components $\mathbf{y_{i,T}}$. Consider a scalar, $y_T = w_{1,T} y_{1,T} + (1 - w_{1,T})\, y_{2,T}$ say. It may be thought that, when the $y_{i,T}$ depend in different ways on the general information set $\mathcal{I}_{\mathcal{T}-\infty}$, predictability could be improved by disaggregation. However, let $\mathsf{E_T}[\mathsf{y_{i,T}}|\mathcal{I}_{\mathcal{T}-\infty}] = \boldsymbol{\delta}'_{\mathbf{i,T}}\mathcal{I}_{\mathcal{T}-\infty}$ then:

$$\mathsf{E_T}[\mathsf{y_T} \mid \mathcal{I}_{\mathcal{T}-\infty}] = \sum_{\mathsf{i}=1}^{2} \mathsf{w_{i,T}} \mathsf{E_T}[\mathsf{y_{i,T}} \mid \mathcal{I}_{\mathcal{T}-\infty}] = \sum_{\mathsf{i}=1}^{2} \mathsf{w_{i,T}} \boldsymbol{\delta}'_{\mathbf{i,T}} \mathcal{I}_{\mathcal{T}-\infty} = \boldsymbol{\lambda}'_{\mathbf{T}} \mathcal{I}_{\mathcal{T}-\infty}$$

say, so nothing is gained unless the previous situation of increased $\mathcal{I}_{\mathcal{T}-\infty}$ is attained. Indeed, if the $w_{i,T}$ change and the $\boldsymbol{\delta}'_{\mathbf{i,T}}$ do not, forecasting the aggregate could well be easier. Thus, the key issue in (say) inflation prediction is not predicting the component price changes, but including those elements in $\mathcal{I}_{\mathcal{T}-\infty}$, rather than restricting $\mathcal{I}_{\mathcal{T}-\infty}$ to lags of aggregate inflation.

2.1.2 Increasing horizon

The obverse of the horizon growing for a fixed event $\mathbf{y_T}$ is that the information set is fixed at $\mathcal{I}_{\mathcal{T}}$ (say), and we consider predictability as the horizon increases for $\mathbf{y_{T+h}}$ as $h = 1, 2, \ldots, H$. If a variable is unpredictable according to (2) (a '1-step' definition), then it must remain unpredictable as the horizon increases $\forall (T+h) \in \mathcal{T}$ (i.e., excluding changes in predictability as considered in the next section): this again follows from (11). Equally, 'looking back' from time $T + h$, the available information sets form a decreasing, nested sequence as in (12). Beyond these rather weak implications, little more can be said in general once densities can change over time. For example, anticipating the next section, consider the non-stationary process:

$$y_t = \rho t + t^{-1}\epsilon_t \;\text{ where }\; \epsilon_t \sim \mathsf{IN}\left[0, \sigma_\epsilon^2\right], \tag{13}$$

where we wish to compare the predictability of y_{T+h} with that of y_{T+h-1} given $\mathcal{I}_T$ for known ρ. Then:

$$\begin{aligned} \mathsf{V}_{\mathsf{T+h}}\left[\mathsf{y}_{\mathsf{T+h}} \mid \mathcal{I}_T\right] &= \mathsf{E}_{\mathsf{T+h}}\left[\left(\mathsf{y}_{\mathsf{T+h}} - \rho\left(\mathsf{T}+\mathsf{h}\right)\right)^2\right] \\ &= \mathsf{E}_{\mathsf{T+h}}\left[\left(\left(\mathsf{T}+\mathsf{h}\right)^{-1}\epsilon_{\mathsf{T+h}}\right)^2\right] \\ &= (T+h)^{-2}\sigma_\epsilon^2 < \mathsf{V}_{\mathsf{T+h-1}}\left[\mathsf{y}_{\mathsf{T+h-1}} \mid \mathcal{I}_T\right]. \end{aligned} \tag{14}$$

The inequality in (14) is strict, and y_{T+h} becomes systematically more predictable from $\mathcal{I}_T$ as h increases. Although DGPs like (13) may be unrealistic, specific assumptions (such as stationarity and ergodicity or mixing) are needed for stronger implications. For example, in a dynamic system which induces error accumulation, where error variances do not decrease systematically as time passes (e.g., being drawn from a mixing process), then predictability falls as the horizon increases since additional unpredictable components will accrue.

2.2 Non-stationarity

In non-stationary processes, unpredictability is also relative to the historical time period considered (which is why the notation above allowed for possibly changing densities), since it is then possible that:

$$\mathsf{D}_{\mathbf{u_t}}\left(\mathbf{u_t} \mid \mathcal{I}_{\sqcup-\infty}\right) \neq \mathsf{D}_{\mathbf{u_t}}\left(\mathbf{u_t}\right) \text{ for } \mathsf{t} = 1,\ldots,\mathsf{T},$$

yet:

$$\mathsf{D}_{\mathbf{u_t}}\left(\mathbf{u_t} \mid \mathcal{I}_{\sqcup-\infty}\right) = \mathsf{D}_{\mathbf{u_t}}\left(\mathbf{u_t}\right) \text{ for } \mathsf{t} = \mathsf{T}+1,\ldots,\mathsf{T}+\mathsf{H},$$

or *vice versa*. More generally, the extent of any degree of predictability can change over time, especially in a social science like economics (e.g., a move from fixed to floating exchange rates).

A major source of non-stationarity in economics derives from the presence of unit roots. However, these can be 'removed' for the purposes of the theoretical analysis by considering suitably differenced or cointegrated combinations of variables, and that is assumed below: section 4 considers the relevant transformations in detail for a vector autoregression. Of course, predictability is thereby changed—a random walk is highly predictable in levels but has unpredictable changes—but it is convenient to consider such I(0) transformations.

In terms of $\mathbf{f_t}\left(\mathcal{I}_{\sqcup-\infty}\right)$ in (6), two important cases of change can now be distinguished. In the first, $\mathbf{f_t}\left(\cdot\right)$ alters to $\mathbf{f_{t+1}}\left(\cdot\right)$, so $\mathbf{f_{t+1}}\left(\cdot\right) \neq \mathbf{f_t}\left(\cdot\right)$, but the resulting mean of the $\{\mathbf{y_t}\}$ process does not change:

$$\mathsf{E}_{\mathsf{t+1}}\left[\mathbf{y_{t+1}}\right] = \mathsf{E}_{\mathsf{t}}\left[\mathbf{y_t}\right]. \tag{15}$$

In the face of such a change, interval predictions may be different, but their mean will be unaltered. In the second case, (15) is violated, so there is a 'location shift' which alters the mean:

$$\mathsf{E}_{\mathsf{t+1}}\left[\mathbf{y_{t+1}}\right] \neq \mathsf{E}_{\mathsf{t}}\left[\mathbf{y_t}\right].$$

Such changes over time are unproblematic for the concept of unpredictability, since $\mathbf{y_{t+j}} - \mathbf{f_{t+j}}\left(\mathcal{I}_{\sqcup+|-\infty}\right)$ is unpredictable for both periods $j = 0, 1$. The practical difficulties, however, for the forecaster may be immense, an issue to which we now turn.

3 Implications for forecasting

It is clear that one cannot forecast the unpredictable beyond its unconditional mean, but there may be hope of forecasting predictable events. To summarize, predictability of a random variable like $\mathbf{y_t}$ in (6) from $\mathcal{I}_{\sqcup-\infty}$ has six distinct aspects:

1. the composition of $\mathcal{I}_{\sqcup-\infty}$;
2. how $\mathcal{I}_{\sqcup-\infty}$ influences $\mathsf{D}_{\mathbf{y_t}}(\cdot \mid \mathcal{I}_{\sqcup-\infty})$ (or specifically, $\mathbf{f_t}(\mathcal{I}_{\sqcup-\infty})$);
3. how $\mathsf{D}_{\mathbf{y_t}}(\cdot \mid \mathcal{I}_{\sqcup-\infty})$ (or specifically $\mathbf{f_t}(\mathcal{I}_{\sqcup-\infty})$) changes over time;
4. the use of the limited information set $\mathcal{J}_{\sqcup-\infty} \subset \mathcal{I}_{\sqcup-\infty}\ \forall t$;
5. the mapping of $\mathsf{D}_{\mathbf{y_t}}(\cdot \mid \mathcal{I}_{\sqcup-\infty})$ into $\mathsf{D}_{\mathbf{y_t}}(\cdot \mid \mathcal{J}_{\sqcup-\infty})$ (or specifically, $\mathbf{f_t}(\mathcal{I}_{\sqcup-\infty})$ into $\mathbf{g_t}(\mathcal{J}_{\sqcup-\infty})$);
6. how $\mathcal{J}_T$ *will* enter $\mathsf{D}_{\mathbf{y_{T+h}}}(\cdot \mid \mathcal{J}_T)$ (or $\mathbf{f_{T+h}}(\mathcal{J}_T)$).

Forecasts of $\mathbf{y_{T+h}}$ from a forecast origin at T are made using the model $\mathbf{y_t} = \boldsymbol{\psi}(\mathcal{J}_{\sqcup-\infty}, \boldsymbol{\theta})$ based on the limited information set $\mathcal{J}_{\sqcup-\infty}$ with conditional expectation $\mathsf{E}[\mathbf{y_t}|\mathcal{J}_{\sqcup-\infty}] = \mathbf{g_t}(\mathcal{J}_{\sqcup-\infty})$. The postulated parameters (or indexes of the assumed distribution) $\boldsymbol{\theta}$ must be estimated as $\widehat{\boldsymbol{\theta}}_T$ using a sample $t = 1, \ldots, T$ of observed information, denoted by $\widehat{\mathcal{J}}_{t-1}$. Doing so therefore introduces four more steps:

7. the approximation of $\mathbf{g_t}(\mathcal{J}_{\sqcup-\infty})$ by a function $\boldsymbol{\psi}(\mathcal{J}_{\sqcup-\infty}, \boldsymbol{\theta})\ \forall t$;
8. measurement errors between $\mathcal{J}_{\sqcup-\infty}$ and the observed $\widehat{\mathcal{J}}_{t-1}\ \forall t$;
9. estimation of $\boldsymbol{\theta}$ in $\boldsymbol{\psi}(\widehat{\mathcal{J}}_{\mathbf{t-1}}, \boldsymbol{\theta})$ from in-sample data $\widehat{\mathcal{J}}_T$;
10. forecasting $\mathbf{y_{T+h}}$ from $\boldsymbol{\psi}_{\mathbf{h}}(\widehat{\mathcal{J}}_{\mathbf{T}}, \widehat{\boldsymbol{\theta}}_{\mathbf{T}})$.

We consider these ten aspects in turn.

Concerning 1., although knowledge of the composition of $\mathcal{I}_{\sqcup-\infty}$ will never be available for such a complicated entity as an economy, any hope of success in forecasting with macro-econometric models requires that they actually do embody inertial responses. Consequently, $\mathcal{I}_{\sqcup-\infty}$ needs to have value for predicting the future evolution of the variables to be forecast, either from a causal or systematic correlational basis. Evidence on this requirement has perforce been based on using $\mathcal{J}_{\sqcup-\infty}$, but seems clear-cut in two areas. First, there is a well-known range of essentially unpredictable financial variables, including changes in exchange rates, E_r, long-term interest rates, R_L, commodity prices P_c and equity prices, P_e: if any of these could be accurately forecast for a future period, a 'money machine' could be created, which in turn would alter the outcome.[1] While these are all key prices in decision taking, forward and future markets have evolved to help offset the risks of changes: unfortunately, there is yet little evidence supporting the efficacy of those markets in forecasting the associated outcomes. Secondly, production processes indubitably take time, so lagged reactions seem the norm on the physical side of the economy. Thus, predictability does not seem to be precluded if $\mathcal{I}_{\sqcup-\infty}$ was known.

Learning precisely how $\mathcal{I}_{\sqcup-\infty}$ is relevant (aspect 2., albeit via $\widehat{\mathcal{J}}_{t-1}$) has been the main focus of macro-econometric modelling, thereby inducing major developments in that discipline, particularly in recent years as various forms of non-stationarity have been modelled. Even so, a lack of well-based empirical equation specifications, past changes in data densities that remain poorly understood, mis-measured—and sometimes missing—data series (especially at frequencies higher than quarterly), and the present limitations of model selection tools to (near) linear models entail that much remains to be achieved at the technical frontier.

Changes in $\mathbf{f_t}(\mathcal{I}_{\sqcup-\infty})$ over time (3.) have been discussed above, and our earlier research has clarified the impacts on forecasting of shifts in its mean values.

[1] A 'fixed-point' analysis (like that proposed by Marget, 1929) is possible, but seems unlikely for phenomena prone to bubbles. However, transactions costs allow some predictability.

Turning to aspect 4., economic theory is the main vehicle for the specification of the information set $\mathcal{J}_{\sqcup-\infty}$, partly supported by empirical studies. Any model of $\mathsf{D}_{\mathbf{y_t}}(\cdot|\cdot)$ embodies $\mathbf{g_t}(\cdot)$ not $\mathbf{f_t}(\cdot)$, but section 2.1 showed that models with mean innovation errors could still be developed. Thus, incomplete information about the 'causal' factors is not by itself problematic, providing $\mathbf{g_t}(\mathcal{J}_{\sqcup-\infty})$ is known.

Unfortunately, mapping $\mathbf{f_t}(\mathcal{I}_{\sqcup-\infty})$ into its conditional expectation $\mathbf{g_t}(\mathcal{J}_{\sqcup-\infty})$ (aspect 5.) is not under the investigator's control beyond the choice of $\mathcal{J}_{\sqcup-\infty}$. Any changes in $\mathbf{f_t}(\mathcal{I}_{\sqcup-\infty})$ over time will have indirect effects on $\mathbf{g_t}(\mathcal{J}_{\sqcup-\infty})$ and make interpreting and modelling these shifts difficult. Nevertheless, the additional mistakes that arise from this mapping act like innovation errors.

However, even if 1.–5. could be overcome in considerable measure, aspect 6. highlights that relationships can change in the future, perhaps dramatically.[2] Section 2.2 distinguished between 'mean-zero' and 'location' shifts in $\mathbf{y_t}$, the most pernicious breaks being location shifts (e.g., confirmed in the forecasting context by the taxonomy of forecast errors in Clements and Hendry, 1998, and by a Monte Carlo in Hendry, 2000). Consider $h = 1$, where the focus is on the mean, $\mathsf{E}_{\mathsf{T}+1}[\mathbf{y_{T+1}}|\mathcal{J}_{\mathcal{T}}]$, which is the integral over the DGP distribution at time $T+1$ conditional on a reduced information set $\mathcal{J}_{\mathcal{T}}$, and hence is unknown at T. Then averaging across alternative choices of the contents of $\mathcal{J}_{\mathcal{T}}$ could provide improved forecasts relative to any single method (i.e., better approximate the integral) when the distribution changes from time T, and those choices reflect different sources of information. Of course, unanticipated breaks that occur after forecasts have been announced cannot be offset: the precise form of $\mathsf{D}_{\mathbf{y_{T+h}}}(\cdot|\cdot)$ is not knowable till time $T+h$ has been reached. However, after time $T+h$, $\mathsf{D}_{\mathbf{y_{T+h}}}(\cdot|\cdot)$ becomes an in-sample density, so thereafter breaks could be offset.

Aspect 7., appears to be the central difficulty: $\mathbf{g_t}(\cdot)$ is not known. First, $\mathbf{g_t}(\mathcal{J}_{\sqcup-\infty})$ experiences derived rather than direct breaks from changes in $\mathbf{f_t}(\mathcal{I}_{\sqcup-\infty})$, making model formulation and especially selection hard. Secondly, empirical modellers perforce approximate $\mathbf{g_t}(\mathcal{J}_{\sqcup-\infty})$ by a function $\boldsymbol{\psi}(\mathcal{J}_{\sqcup-\infty}, \boldsymbol{\theta})$, where the formulation of $\boldsymbol{\theta}$ is intended to incorporate the effects of past breaks: most 'time-varying coefficient', regime-switching, and non-linear models are members of this class. Thirdly, while 'modelling breaks' may be possible for historical events, a location shift at, or very near, the forecast origin may not be known to the forecaster; and even if known, may have effects that are difficult to discern, and impossible to model with the limited information available.

Measurement errors, aspect 8., almost always arise, as available observations are inevitably inaccurate. Although these may bias estimated coefficients and compound the modelling difficulties, by themselves, measurement errors do not imply inaccurate forecasts relative to the measured outcomes. However, in dynamic models, measurement errors induce negative moving-average residuals. Thus, a potential incompatibility arises: differencing to attentuate systematic mis-specification or a location shift will exacerbate a negative moving average. Conversely, a forecast-error correction can remove unit roots and hence lose robustness to breaks. This new result seems to lie at the heart of practical forecasting problems, and may explain the many cases where (e.g.) differencing and intercept corrections have performed badly.

Concerning aspect 9., the 'averaging' of historical data to estimate $\boldsymbol{\theta}$ by $\widehat{\boldsymbol{\theta}}_T$ imparts additional inertia in the model relative to the data, as well as increased uncertainty. More importantly, there are probably estimation biases from not fully capturing all past breaks, which would affect deterministic terms.

Finally, concerning aspect 10., multistep forecasts have the added difficulty of cumulative errors although these are no more than would arise in the context of predictability.

[2] Sir Alec Cairncross (1969) suggested the example of forecasting UK GNP in 1940 for 1941—a completely different outcome would have materialized had an invasion occurred. The recent theoretical analyses discussed above have in fact helped to formalize many of the issues he raised.

Not adapting to location breaks induces systematic mis-forecasting, usually resulting in forecast failure. To thrive competitively, forecasting models need to avoid that fate, as there are many devices that track (with a lag) and hence are robust to such breaks once they have occurred. Section 5 considers several such devices. Before that, however, sub-section 3.1 formalizes these possible errors in a taxonomy to seek pointers for attenuation of their adverse consequences.

3.1 Taxonomy of error sources

To forecast $\mathbf{y_{T+h}}$, the in-sample model $\boldsymbol{\psi}(\widehat{\mathcal{J}}_{\mathbf{T}}, \widehat{\boldsymbol{\theta}}_{\mathbf{T}})$ is developed for some specification of the parameters $\boldsymbol{\theta} \in \mathbb{R}$ estimated as $\widehat{\boldsymbol{\theta}}_T$ from the full-sample information $\widehat{\mathcal{J}}_T$ where $\mathcal{J}_{\sqcup-\infty} \subseteq \mathcal{I}_{\sqcup-\infty}$ is the available information set at each point in time, measured by $\widehat{\mathcal{J}}_{t-1}$ such that:

$$\widehat{\mathbf{y}}_{T+h|T} = \boldsymbol{\psi}_{\mathbf{h}}\left(\widehat{\mathcal{J}}_{\mathbf{T}}, \widehat{\boldsymbol{\theta}}_{\mathbf{T}}\right). \tag{16}$$

There are many ways to formulate the function $\boldsymbol{\psi}_{\mathbf{h}}(\cdot)$ in (16) for a dynamic model $\boldsymbol{\psi}(\cdot)$, including 'powering up' and multi-step estimation. Below, only the former is considered (on the latter, see Bhansali, 1996, 1997, 1999, Clements and Hendry, 1996, and Chevillon and Hendry, 2002, *inter alia*), but this section allows for any possibility. Conversely, we focus on the first two moments here rather than the complete forecast distribution.

The key steps that determine the forecast error:

$$\widehat{\mathbf{u}}_{T+h|T} = \mathbf{y}_{\mathbf{T+h}} - \widehat{\mathbf{y}}_{\mathbf{T+h|T}} = \mathbf{f}_{\mathbf{T+h}}\left(\mathcal{I}_{\mathcal{T}+\langle-\infty}\right) + \boldsymbol{\nu}_{\mathbf{T+h}} - \boldsymbol{\psi}_{\mathbf{h}}(\widehat{\mathcal{J}}_{\mathbf{T}}, \widehat{\boldsymbol{\theta}}_{\mathbf{T}}),$$

are: the composition of the DGP information sets $\mathcal{I}_{\sqcup-\infty}$; how each $\mathcal{I}_{\sqcup-\infty}$ enters the DGP $\mathsf{D}_{\mathbf{y_t}}(\mathbf{y_t}|\mathcal{I}_{\sqcup-\infty})$; how $\mathsf{D}_{\mathbf{y_t}}(\mathbf{y_t}|\mathcal{I}_{\sqcup-\infty})$ changes over time in-sample; the limited information set $\mathcal{J}_{\sqcup-\infty} \subseteq \mathcal{I}_{\sqcup-\infty}$; the mapping of $\mathsf{D}_{\mathbf{y_t}}(\mathbf{y_t}|\mathcal{I}_{\sqcup-\infty})$ into $\mathsf{D}_{\mathbf{y_t}}(\mathbf{y_t}|\mathcal{J}_{\sqcup-\infty})$ inducing $\mathbf{g_t}(\mathcal{J}_{\sqcup-\infty}) = \mathsf{E}_{\mathbf{t}}[\mathbf{f_t}(\mathcal{I}_{\sqcup-\infty})|\mathcal{J}_{\sqcup-\infty}]$; how $\mathcal{J}_{\mathcal{T}}$ will enter $\mathsf{D}_{\mathbf{y_{T+h}}}(\cdot|\mathcal{J}_{\mathcal{T}})$ for a forecast origin at T; the approximation of $\mathbf{g_t}(\mathcal{J}_{\sqcup-\infty})$ by the model $\boldsymbol{\psi}(\mathcal{J}_{\sqcup-\infty}, \boldsymbol{\theta})$; the specification of $\boldsymbol{\theta}$; measurement errors in each $\widehat{\mathcal{J}}_{t-1}$ for $\mathcal{J}_{\sqcup-\infty}$ (which may themselves change over time); and the estimation of $\boldsymbol{\theta}$ by $\widehat{\boldsymbol{\theta}}_T$, which together determine the properties of $\boldsymbol{\psi}_{\mathbf{h}}(\cdot)$. The first six are aspects of predictability in the DGP; the second four of the formulation of forecasting models which seek to capture that predictability.

From such a formulation, $\widehat{\mathbf{u}}_{T+h|T}$ can be decomposed into errors which derive from each of the main reduction or transformation steps, namely:

$$\begin{aligned}\widehat{\mathbf{u}}_{T+h|T} \;=\; & \boldsymbol{\nu}_{\mathbf{T+h}} + \left[\mathbf{f}_{\mathbf{T+h}}\left(\mathcal{I}_{\mathcal{T}+\langle-\infty}\right) - \mathbf{f}_{\mathbf{T+h}}(\mathcal{I}_{\mathcal{T}})\right] + \left[\mathbf{f}_{\mathbf{T+h}}(\mathcal{I}_{\mathcal{T}}) - \mathbf{g}_{\mathbf{T+h}}(\mathcal{J}_{\mathcal{T}})\right] + \left[\mathbf{g}_{\mathbf{T+h}}(\mathcal{J}_{\mathcal{T}}) - \mathbf{g}_{\mathbf{T+h|T}}(\mathcal{J}_{\mathcal{T}})\right] \\ & + \left[\mathbf{g}_{\mathbf{T+h|T}}(\mathcal{J}_{\mathcal{T}}) - \boldsymbol{\psi}_{\mathbf{h}}(\mathcal{J}_{\mathcal{T}}, \boldsymbol{\theta})\right] + \left[\boldsymbol{\psi}_{\mathbf{h}}(\mathcal{J}_{\mathcal{T}}, \boldsymbol{\theta}) - \boldsymbol{\psi}_{\mathbf{h}}(\widehat{\mathcal{J}}_{\mathbf{T}}, \boldsymbol{\theta})\right] + \left[\boldsymbol{\psi}_{\mathbf{h}}(\widehat{\mathcal{J}}_{\mathbf{T}}, \boldsymbol{\theta}) - \boldsymbol{\psi}_{\mathbf{h}}(\widehat{\mathcal{J}}_{\mathbf{T}}, \widehat{\boldsymbol{\theta}}_{\mathbf{T}})\right]\end{aligned} \tag{17}$$

where $\mathbf{g}_{\mathbf{T+h|T}}(\mathcal{J}_{\mathcal{T}})$ is the 'extrapolated' value of $\mathbf{g}_{\mathbf{T+h}}(\mathcal{J}_{\mathcal{T}})$ for constant forecast-origin parameters in $\mathbf{g}(\cdot)$. While decompositions such as (17) are not unique, they help pinpoint the potential sources of forecast failure, and which components are less likely to have a pernicious effect on forecast accuracy.

Taking the seven right-hand side terms in (17) in turn, the first four are unknowable (in the absence of a crystal ball), being dependent on the future innovation $\boldsymbol{\nu}_{\mathbf{T+h}}$, future information accrual, the change to the limited information set, and post-forecast-origin changes in the induced process: all 4 are, therefore, unpredictable, will affect the forecast-error variance, and may influence its mean. The first, second and third terms have expected values of zero for proper information sets $\mathcal{I}$ and $\mathcal{J}$, so will not affect $\mathsf{E}_{\mathsf{T+h}}\left[\widehat{\mathbf{u}}_{\mathsf{T+h|T}}|\mathcal{J}_{\mathcal{T}}\right]$. Consequently, a lack of knowledge of the complete information set $\mathcal{I}$ is not an explanation for forecast failure, a general result of importance below, although using more (relevant) information will reduce the variance component from $\mathbf{f}_{\mathbf{T+h}}(\mathcal{I}_{\mathcal{T}}) - \mathbf{g}_{\mathbf{T+h}}(\mathcal{J}_{\mathcal{T}})$. The

second term is only present when $h > 1$, but then represents the cumulation of the innovation errors $\{\boldsymbol{\nu}_{\mathbf{T+j}}\}$ for $j = 1, \ldots, h-1$. However, the fourth term is a potential source of forecast failure when $\mathbf{g}_{\mathbf{T+h}}(\mathcal{J}_{\mathcal{T}}) \neq \mathbf{g}_{\mathbf{T+h|T}}(\mathcal{J}_{\mathcal{T}})$. That requires an induced location shift to be non-zero on average, rather than just structural change in general. Conversely, the third term would be zero under constant parameters.

The next three terms depend on the goodness of the model for the local DGP $\mathsf{D}_{\mathbf{y_T}}(\mathbf{y_T}|\mathcal{J}_{\mathcal{T}})$ and on data accuracy, both in-sample and at the forecast origin, as well as the choice of estimator. Specifically, the fifth is a function of the adequacy of the model, the sixth of the data accuracy at T, and the last on the properties of the estimator $\widehat{\boldsymbol{\theta}}_T$ for $\boldsymbol{\theta}$ when the observed data are used. Thus, the fifth term would be zero for a correctly specified model, the sixth for accurate data, but the seventh only in an infinite sample, hence the focus in many derivations of forecast-error uncertainties on the impacts of parameter estimation and innovation error variances.

The 1-step ahead error from the forecasting model $\widehat{\mathbf{y}}_{T+1} = \boldsymbol{\psi}_{\mathbf{1}}(\widehat{\mathcal{J}}_{\mathbf{T}}, \widehat{\boldsymbol{\theta}}_{\mathbf{T}})$ is $\mathbf{u_{T+1}} = \mathbf{y_{T+1}} - \widehat{\mathbf{y}}_{\mathbf{T+1}}$. Then $\mathbf{u_{T+1}}$ can be decomposed into six basic sources of mistakes (as can further-ahead errors):

$\mathbf{u_{T+1}}$	$=$	$\boldsymbol{\nu}_{\mathbf{T+1}}$	DGP innovation error
	$+$	$\mathbf{f}_{\mathbf{T+1}}(\mathcal{I}_{\mathcal{T}}) - \mathbf{g}_{\mathbf{T+1}}(\mathcal{J}_{\mathcal{T}})$	incomplete information
	$+$	$\mathbf{g}_{\mathbf{T+1}}(\mathcal{J}_{\mathcal{T}}) - \mathbf{g}_{\mathbf{T}}(\mathcal{J}_{\mathcal{T}})$	induced change
	$+$	$\mathbf{g}_{\mathbf{T}}(\mathcal{J}_{\mathcal{T}}) - \boldsymbol{\psi}_{\mathbf{1}}(\mathcal{J}_{\mathcal{T}}, \boldsymbol{\theta})$	approximation reduction
	$+$	$\boldsymbol{\psi}_{\mathbf{1}}(\mathcal{J}_{\mathcal{T}}, \boldsymbol{\theta}) - \boldsymbol{\psi}_{\mathbf{1}}(\widehat{\mathcal{J}}_{\mathbf{T}}, \boldsymbol{\theta})$	measurement error
	$+$	$\boldsymbol{\psi}_{\mathbf{1}}\left(\widehat{\mathcal{J}}_{\mathbf{T}}, \boldsymbol{\theta}\right) - \boldsymbol{\psi}_{\mathbf{1}}\left(\widehat{\mathcal{J}}_{\mathbf{T}}, \widehat{\boldsymbol{\theta}}_{\mathbf{T}}\right)$	estimation uncertainty

We consider these in turn.

Since $\boldsymbol{\nu}_{\mathbf{T+1}}$ is an innovation against the DGP information set $\mathcal{I}_{\mathcal{T}}$, nothing will reduce its uncertainty. Nevertheless, the intrinsic properties of $\boldsymbol{\nu}_{\mathbf{T+1}}$ matter greatly, specifically its variance, and any unpredictable changes in its distribution. The baseline accuracy of a forecast cannot exceed that inherited from the DGP innovation error.

There are many reasons why information available to the forecaster is incomplete relative to that underlying the behaviour of the DGP. For example, important variables may not be known, and even if known, may not be measured. Either of these make $\mathcal{J}_{\mathcal{T}}$ a subset of $\mathcal{I}_{\mathcal{T}}$, although the first (excluding relevant information) tends to be the most emphasised. As shown in section 2.1, incomplete information increases forecast uncertainty over any inherent unpredictability, but by construction:

$$\mathbf{g}_{\mathbf{T+1}}(\mathcal{J}_{\mathcal{T}}) = \mathsf{E}_{\mathsf{T+1}}[\mathbf{f}_{\mathbf{T+1}}(\mathcal{I}_{\mathcal{T}}) \mid \mathcal{J}_{\mathcal{T}}],$$

so, no additional biases result from this source, even when breaks often occur.

Rather, the problems posed by breaks manifest themselves in the next term, $\mathbf{g}_{\mathbf{T+1}}(\mathcal{J}_{\mathcal{T}}) - \mathbf{g}_{\mathbf{T}}(\mathcal{J}_{\mathcal{T}})$: sub-section 3.3 below addresses their detection. In-sample, it is often possible to ascertain that a break has occurred, and at worst develop suitable indicator variables to offset it, but the real difficulties derive from breaks at, or very near, the forecast origin. Sub-section 3.4 considers possible remedies: here we note that if $\Delta\mathbf{g}_{\mathbf{T+1}}(\mathcal{J}_{\mathcal{T}})$ has a non-zero mean, either an additional intercept (i.e., intercept correction, denoted IC), or further differencing will remove that mean error.

There will also usually be mis-specifications due to the formulation of both $\boldsymbol{\psi}(\cdot)$ and $\boldsymbol{\theta}$ as approximations to $\mathbf{g}_{\mathbf{T}}(\mathcal{J}_{\mathcal{T}})$. For example, linear approximations to non-linear responses will show up here, as will dynamic mis-specification ($\mathcal{J}_{\mathcal{T}}$ assumes all earlier values are available, but models often impose short lag lengths). If the effect is systematic, then an IC or differencing will again reduce its impact; however the required sign may be incompatible with the previous case.

Even if all variables known to be relevant are measured, the observations available may be inaccurate relative to the DGP 'forces'. A distinction from the case of excluding relevant information is useful, as it matters what the source is: measurement errors in dynamic models tend to induce negative moving average residuals, whereas omitted variables usually lead to positive autoregressive residuals. Thus, again a potential incompatibility arises: differencing will exacerbate a negative moving average, and an IC may need the opposite sign to that for a break.

Finally, estimation uncertainty arising from using $\widehat{\boldsymbol{\theta}}_T$ in place of $\boldsymbol{\theta}$ can compound the systematic effects of breaks when $\widehat{\boldsymbol{\theta}}_T$ adjusts slowly to changes induced in $\boldsymbol{\theta}$.

When models are mis-specified by using $\mathcal{J}_{\sqcup-\infty} \subset \mathcal{I}_{\sqcup-\infty}$, for a world where $\mathcal{I}_{\sqcup-\infty}$ enters the density in changing ways over time, forecasting theory delivers implications that are remarkably different from the theorems that hold for constant processes as the summary discussion in Hendry and Clements (2003) emphasises. We can now see a basis for such results in the gulf between predictability and empirical forecasting highlighted by the above taxonomy.

3.2 Congruent modelling for forecasting

Given the taxonomy, what is role for orthogonalised, parsimonious encompassing, congruent models? Eight benefits are potentially available, even in the forecasting context, and the need for such a model in the policy context is clear.

1. Rigorous in-sample modelling helps detect and thereby avoid equilibrium-mean shifts which would otherwise distort forecasts.
2. Such models deliver the smallest variance for the innovation error defined on the available information set, and hence offer one measure of the 'best approximation' to $g(\cdot)$.
3. It is important to remove irrelevant variables which might suffer breaks over the forecast horizon (see e.g., Clements and Hendry, 2002).
4. The best estimates of the model's parameters will be invaluable over periods when no breaks occur, and thereby reduce forecast-error variances.
5. An orthogonalised and parsimonious model will avoid a large ratio of the largest to smallest eigenvalue of the second-moment matrix, which can have a detrimental effect on forecast-error variances when second moments alter, even for constant parameters in the forecasting model.
6. A dominant parsimonious congruent model offers better understanding of the economic process by being more interpretable.
7. Such a model also sustains a progressive research strategy and offers a framework for interpreting forecast failure.

Nevertheless, how such a model is used in the forecast period also matters and is discussed below.

3.3 Diagnosing breaks

A problem for the forecaster hidden in the above formulation is determining that there has been a break. First, data at or near the forecast origin are always less well measured than more mature vintages, and some may be missing. Thus, a recent forecast error may reflect just a data mistake, and treating it as a location shift in the economy could induce systematic forecast errors in later periods. Secondly, a model which is mis-specified for the underlying process, such as a linear autoregression fitted to a regime-switching DGP, may suggest breaks have occurred when they have not. Then, 'solutions' such as additional differencing or intercept corrections (ICs) need not be appropriate. Thirdly, even when a break has occurred in some part of a model, its effects elsewhere depend on how well both the relevant equations and their links are specified: UK M1 below provides an example where only the

opportunity cost is mis-forecast in one version of the model, but real money is in another. Fourthly, sudden changes to data (e.g., in observed money growth rates) need not entail a break in the associated equation of the model: UK M1 again highlights this. Thus, without knowing how well specified a model is under recently changed conditions, data movements alone are insufficient to guide the detection of breaks. Unfortunately, therefore, only recent forecast errors are useful for diagnosing change relative to a model, highlighting the importance of distinguishing additive from innovation errors.

3.3.1 Co-breaking

On the other hand, co-breaking of a subset of relations over the forecast horizon would be valuable because such variables would move in tandem as a group. Although forecasting the remaining variables would still be problematic, one would not need ICs for the co-breaking equations, which would improve the efficiency of the forecasts. The UK M1 system also illustrates this aspect, as an IC is needed in only one equation.

Moreover, lagged co-breaking is invaluable. A break in a marginal process, which affects the variable to be forecast with a lag, does not induce forecast failure.

3.4 Potential improvements

A reduction in the seriousness of forecast failure could be achieved by:

(a) breaks being sufficiently infrequent to ignore;
(b) a forecasting system being invariant to breaks;
(c) an investigator forecasting breaks; or
(d) forecasts adapting rapidly to breaks that occur.

All four possibilities merit consideration.[3]

(a) relates to the second role of data frequency noted above. If breaks occur erratically over time and across variables, but with an average of once per r years per variable (where r could be less than unity, but seems larger in practice) then on (e.g.) weekly data, breaks occur once per $52r$ observations. While the impact of any break in a dynamic system takes time to reach its full effect, and high-frequency data are often noisy, nevertheless on such data there will be many periods of 'normal' behaviour between breaks during which 'causal' models should perform well (assuming past breaks have been successfully modelled). Conversely, breaks will be relatively frequent on annual data (roughly 15% of the time for GDP since 1880 in the UK: see Clements and Hendry, 2001b). Analyses of other series for breaks to ascertain their size and latency distributions would be useful, perhaps using robust univariate devices as the baseline against which to determine the existence and timing of breaks.

When the 'target' variable $\mathbf{y_{T+1}}$ to be forecast is, say, annual inflation, then 'solution' (a) is infeasible: that selection entails the choice of data frequency. However, the frequency need not be the same for $\mathcal{J}_T$: forecasting annual changes from quarterly data is common. Since predictability cannot fall with a larger information set, an implication is to use the highest frequency, and the largest set, irrespective of the 'target' (e.g., hourly data even if annual GNP growth is to be forecast). Although this is usually impractical given the limited sample periods available in macro-economics, and the lack of collection of high-frequency data on many variables of interest, that implication also merits exploration.

(b) unfortunately seems unlikely, and has not happened historically. But it is important to clarify the reason why (b) is unlikely to occur. It is not because autonomous equations are necessarily scarce, but

[3] Averaging a set of forecasts is shown in Hendry and Clements (2004) to improve forecasting when at least one (different) method responds to each break.

because the weakest link in the system determines the overall outcome. For example, consider the oil crisis in the mid 1970s: models which excluded oil prices would certainly have mis-forecast inflation, and experienced 'breaks'—but even models with oil prices would have suffered forecast failure unless they could have forecast the oil crisis itself. After the event, however, a distinction emerges: the former would still suffer serious mis-fitting (probably adapted to by changes in estimated coefficients given the propensity to use least squares estimation which seeks to reduce the largest errors), whereas the latter would not for the inflation equation, but still would for its oil price equation. 'Explaining' the latter by building a model of oil supply would push the problem down a layer, but at some stage, an unanticipated jump is left: a non-linear process–or even an indicator–would remove the misfit *ex post*, but neither need help to forecast the next jump.

(c) essentially requires a crystal ball that can foresee looming changes. In some cases, however, this may be possible. For example, related situations may have occurred previously, allowing a model to be built of the 'change' process itself (though that too could change): regime-switching models are one attempt to do so for states that often change and are partly predictable as the conditional probability of the state differs from the unconditional. To date, their forecasting performance has not proved spectacular, even against univariate predictors, partly because the timing of the switch remains somewhat elusive—albeit crucial to their accuracy. Another possibility is that although breaks are relatively rare, they have discernible precursors, either leading indicators or causal, as is being discovered in volcanology. Here, more detailed studies of evolving breaks are merited.

(d) is more easily implemented, as there many forecasting devices that are robust to various forms of break. Notice the key difference from (c): here adaptability is after the event, improving *ex post* tracking and thereby avoiding systematic forecast failure, whereas (c) sought to improve predictability. As emphasized by Clements and Hendry (1998, 1999), knowing in-sample causal relations need not deliver 'better' forecasts (on some measures) than those from devices where no causal variables are used. Thus, it seems crucial to embed macro-econometric models in a forecasting strategy, where progressive research is essential to unravel (b) and (c), and adaptability after shifts is the key to mitigating (d).

4 A cointegrated DGP

Consider a first-order VAR for simplicity, where the vector of n variables of interest is denoted by $\mathbf{x_t}$, and its DGP is:

$$\mathbf{x_t} = \boldsymbol{\tau} + \boldsymbol{\Upsilon}\mathbf{x_{t-1}} + \boldsymbol{\epsilon_t} \text{ where } \boldsymbol{\epsilon_t} \sim \mathsf{IN_n}\left[\mathbf{0}, \boldsymbol{\Omega}\right]. \tag{18}$$

$\boldsymbol{\Upsilon}$ is an $n \times n$ matrix of coefficients and $\boldsymbol{\tau}$ is an n dimensional vector of constant terms. The specification in (18) is assumed constant in-sample, and the system is taken to be $\mathsf{I}(1)$, satisfying the $r < n$ cointegration relations:

$$\boldsymbol{\Upsilon} = \mathbf{I_n} + \boldsymbol{\alpha}\boldsymbol{\beta}'. \tag{19}$$

$\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are $n \times r$ full-rank matrices, no roots of $|\mathbf{I} - \boldsymbol{\Upsilon}\mathbf{L}| = 0$ lie inside unit circle ($L^k x_t = x_{t-k}$), and $\boldsymbol{\alpha}'_\perp \boldsymbol{\Upsilon} \boldsymbol{\beta}_\perp$ is full rank, where $\boldsymbol{\alpha}_\perp$ and $\boldsymbol{\beta}_\perp$ are full column rank $n \times (n-r)$ matrices, with $\boldsymbol{\alpha}'\boldsymbol{\alpha}_\perp = \boldsymbol{\beta}'\boldsymbol{\beta}_\perp = \mathbf{0}$. Then (18) is reparameterized as a vector equilibrium-correction model (VEqCM):

$$\Delta\mathbf{x_t} = \boldsymbol{\tau} + \boldsymbol{\alpha}\boldsymbol{\beta}'\mathbf{x_{t-1}} + \boldsymbol{\epsilon_t}. \tag{20}$$

Both $\Delta\mathbf{x_t}$ and $\boldsymbol{\beta}'\mathbf{x_t}$ are $\mathsf{I}(0)$ but may have non-zero means. Let:

$$\boldsymbol{\tau} = \boldsymbol{\gamma} - \boldsymbol{\alpha}\boldsymbol{\mu} \tag{21}$$

then:

$$(\Delta \mathbf{x_t} - \boldsymbol{\gamma}) = \boldsymbol{\alpha}\left(\boldsymbol{\beta}'\mathbf{x_{t-1}} - \boldsymbol{\mu}\right) + \boldsymbol{\epsilon_t}. \tag{22}$$

The variables grow at the rate $\mathsf{E}[\Delta \mathbf{x_t}] = \boldsymbol{\gamma}$ with $\boldsymbol{\beta}'\boldsymbol{\gamma} = \mathbf{0}$; and when $\boldsymbol{\beta}'\boldsymbol{\alpha}$ is non-singular, the long-run equilibrium is:

$$\mathsf{E}\left[\boldsymbol{\beta}'\mathbf{x_t}\right] = \boldsymbol{\mu}. \tag{23}$$

Thus, in (22), both $\Delta \mathbf{x_t}$ and $\boldsymbol{\beta}'\mathbf{x_t}$ are expressed as deviations about their means. Note that $\boldsymbol{\gamma}$ is $n \times 1$ subject to r restrictions, and $\boldsymbol{\mu}$ is $r \times 1$, leaving n unrestricted intercepts in total. Also, $\boldsymbol{\gamma}$, $\boldsymbol{\alpha}$ and $\boldsymbol{\mu}$ are assumed to be variation free, although in principle, $\boldsymbol{\mu}$ could depend on $\boldsymbol{\gamma}$: see Hendry and von Ungern-Sternberg (1981). Then $(\boldsymbol{\tau}, \boldsymbol{\Upsilon})$ are not variation free, as seems reasonable when $\boldsymbol{\gamma}$, $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ and $\boldsymbol{\mu}$ are the 'deep' parameters: for a more extensive analysis, see Clements and Hendry (1996).

4.1 Location shifts

The shift of interest here is $\nabla\boldsymbol{\mu}^* = \boldsymbol{\mu}^* - \boldsymbol{\mu}$. Then:

$$\Delta \mathbf{x_{T+1}} = \boldsymbol{\gamma} + \boldsymbol{\alpha}\left(\boldsymbol{\beta}'\mathbf{x_T} - \boldsymbol{\mu}^*\right) + \boldsymbol{\epsilon_{T+1}} \tag{24}$$

so from (24):

$$\Delta \mathbf{x_{T+1}} = \boldsymbol{\gamma} + \boldsymbol{\alpha}\left(\boldsymbol{\beta}'\mathbf{x_T} - \boldsymbol{\mu}\right) + \boldsymbol{\epsilon_{T+1}} - \boldsymbol{\alpha}\boldsymbol{\nabla}\boldsymbol{\mu}^* \tag{25}$$

or:

$$\Delta \mathbf{x_{T+1}} = \widehat{\boldsymbol{\Delta}\mathbf{x}}_{\mathbf{T+1|T}} - \boldsymbol{\alpha}\boldsymbol{\nabla}\boldsymbol{\mu}^*. \tag{26}$$

The first right-hand side term in (26) (namely $\widehat{\Delta \mathbf{x}}_{T+1|T}$) is the constant-parameter forecast of $\Delta \mathbf{x_{T+1}}$; the second is the shift with:

$$\mathsf{E}\left[\Delta \mathbf{x_{T+1}} - \widehat{\boldsymbol{\Delta}\mathbf{x}}_{\mathbf{T+1|T}}\right] = -\boldsymbol{\alpha}\boldsymbol{\nabla}\boldsymbol{\mu}^*.$$

Section 5 now considers possible solutions to avoiding forecast failure.

5 Adaptive devices

Three approaches to implementing suggestion (d) in section 3.4 are considered:

- differencing the VEqCM (22) to improve its forecasting robustness to location shifts;
- rapid updating of the estimates of $\boldsymbol{\gamma}$ and $\boldsymbol{\mu}$ after such shifts; and
- forecast-error corrections to adjust quickly to breaks.

We take these in turn: none actually alters predictability (as the information set is unchanged), but they all seek to mitigate the impact of breaks.

5.1 Differencing the VEqCM

Since shifts in $\boldsymbol{\mu}$ are the most pernicious for forecasting, consider forecasting not from (22) itself but from a variant thereof which has been differenced after a congruent representation has been estimated:

$$\Delta \mathbf{x_t} = \boldsymbol{\Delta}\mathbf{x_{t-1}} + \boldsymbol{\alpha}\boldsymbol{\beta}'\boldsymbol{\Delta}\mathbf{x_{t-1}} + \boldsymbol{\Delta}\boldsymbol{\epsilon_t} = \left(\mathbf{I_n} + \boldsymbol{\alpha}\boldsymbol{\beta}'\right)\boldsymbol{\Delta}\mathbf{x_{t-1}} + \mathbf{u_t} \tag{27}$$

or:

$$\Delta^2 \mathbf{x_t} = \boldsymbol{\alpha}\boldsymbol{\beta}'\boldsymbol{\Delta}\mathbf{x_{t-1}} + \mathbf{u_t}. \tag{28}$$

(27) is just the first difference of the original VAR, since $\left(\mathbf{I_n} + \boldsymbol{\alpha}\boldsymbol{\beta}'\right) = \boldsymbol{\Upsilon}$, but with the rank restriction from cointegration imposed. The alternative representation in (28) can be interpreted as augmenting a double differenced VAR (DDV) forecast by $\boldsymbol{\alpha}\boldsymbol{\beta}'\boldsymbol{\Delta}\mathbf{x_{t-1}}$, which is zero on average.

To trace the behaviour of (27) after a break in $\boldsymbol{\mu}$, let:

$$\widetilde{\Delta \mathbf{x}}_{T+1|T} = \left(\mathbf{I_n} + \boldsymbol{\alpha}\boldsymbol{\beta}'\right) \Delta \mathbf{x_T},$$

where from (25):

$$\Delta \mathbf{x_{T+1}} = \boldsymbol{\Delta} \mathbf{x_T} + \boldsymbol{\alpha} \left(\boldsymbol{\beta}' \boldsymbol{\Delta} \mathbf{x_T} - \boldsymbol{\Delta} \boldsymbol{\mu}^*\right) + \boldsymbol{\Delta} \boldsymbol{\epsilon_{T+1}}.$$

At time T only, $\Delta \boldsymbol{\mu}^* = \boldsymbol{\nabla} \boldsymbol{\mu}^*$, so:

$$\Delta \mathbf{x_{T+1}} = \boldsymbol{\Delta} \mathbf{x_T} + \boldsymbol{\alpha}\boldsymbol{\beta}' \boldsymbol{\Delta} \mathbf{x_T} - \boldsymbol{\alpha} \boldsymbol{\nabla} \boldsymbol{\mu}^* + \boldsymbol{\Delta} \boldsymbol{\epsilon_{T+1}}.$$

Then:

$$\mathsf{E}\left[\Delta \mathbf{x_{T+1}} - \widetilde{\boldsymbol{\Delta} \mathbf{x}}_{\mathbf{T+1|T}}\right] = \Delta \mathbf{x_T} + \boldsymbol{\alpha}\boldsymbol{\beta}' \boldsymbol{\Delta} \mathbf{x_T} - \boldsymbol{\alpha} \boldsymbol{\nabla} \boldsymbol{\mu}^* - \left(\mathbf{I_n} + \boldsymbol{\alpha}\boldsymbol{\beta}'\right) \boldsymbol{\Delta} \mathbf{x_T} = -\boldsymbol{\alpha} \boldsymbol{\nabla} \boldsymbol{\mu}^*.$$

Here there is no gain, as the break is after forecasts are announced—an IC, or DDV, would fare no better.

However, one period later:

$$\Delta \mathbf{x_{T+2}} = \boldsymbol{\Delta} \mathbf{x_{T+1}} + \boldsymbol{\alpha} \left(\boldsymbol{\beta}' \boldsymbol{\Delta} \mathbf{x_{T+1}} - \boldsymbol{\Delta} \boldsymbol{\mu}^*\right) + \boldsymbol{\Delta} \boldsymbol{\epsilon_{T+2}},$$

and now $\Delta \boldsymbol{\mu}^* = \mathbf{0}$, so:

$$\mathsf{E}\left[\Delta \mathbf{x_{T+2}} - \widetilde{\boldsymbol{\Delta} \mathbf{x}}_{\mathbf{T+2|T+1}}\right] = \mathsf{E}\left[\Delta \mathbf{x_{T+1}} + \boldsymbol{\alpha}\boldsymbol{\beta}' \boldsymbol{\Delta} \mathbf{x_{T+1}} - \left(\mathbf{I_n} + \boldsymbol{\alpha}\boldsymbol{\beta}'\right) \boldsymbol{\Delta} \mathbf{x_{T+1}}\right] = \mathbf{0}.$$

Thus, the differenced VEqCM 'misses' for 1 period only, and does not make systematic, and increasing, errors. The next sub-section considers the impact of unnecessary differencing on forecast-error variances, and in the context of 1-step ahead forecasts.

5.1.1 Forecast-error variances

Let $\mathbf{e_{T+1}} = \boldsymbol{\Delta} \mathbf{x_{T+1}} - \widetilde{\boldsymbol{\Delta} \mathbf{x}}_{\mathbf{T+1|T}}$ be the forecast error, then, ignoring parameter estimation uncertainty as $\mathsf{O}_p\left(T^{-1/2}\right)$:

$$\mathbf{e_{T+1}} = -\boldsymbol{\alpha} \boldsymbol{\nabla} \boldsymbol{\mu}^* + \boldsymbol{\Delta} \boldsymbol{\epsilon_{T+1}},$$

and:

$$\mathbf{e_{T+2}} = \boldsymbol{\Delta} \boldsymbol{\epsilon_{T+2}}.$$

Since the system error is $\{\boldsymbol{\epsilon_t}\}$, then the additional differencing doubles the 1-step error variance, which is the same as for the DDV. Relative to a DDV, however, there is a gain from the DVEqCM, since the former has a component from the variance of the omitted variable ($\boldsymbol{\alpha}\boldsymbol{\beta}'\boldsymbol{\Delta}\mathbf{x_t}$), as well as the same error terms. Thus, a DDV is not only the difference of a DVAR, but is also obtained by dropping a mean-zero term from the differenced VEqCM.

Using $\Delta \mathbf{x_T}$ to forecast

Second differencing removes two unit roots, any intercepts and linear trends, changes location shifts to 'blips', and converts breaks in trends to impulses. Figure 1 illustrates.

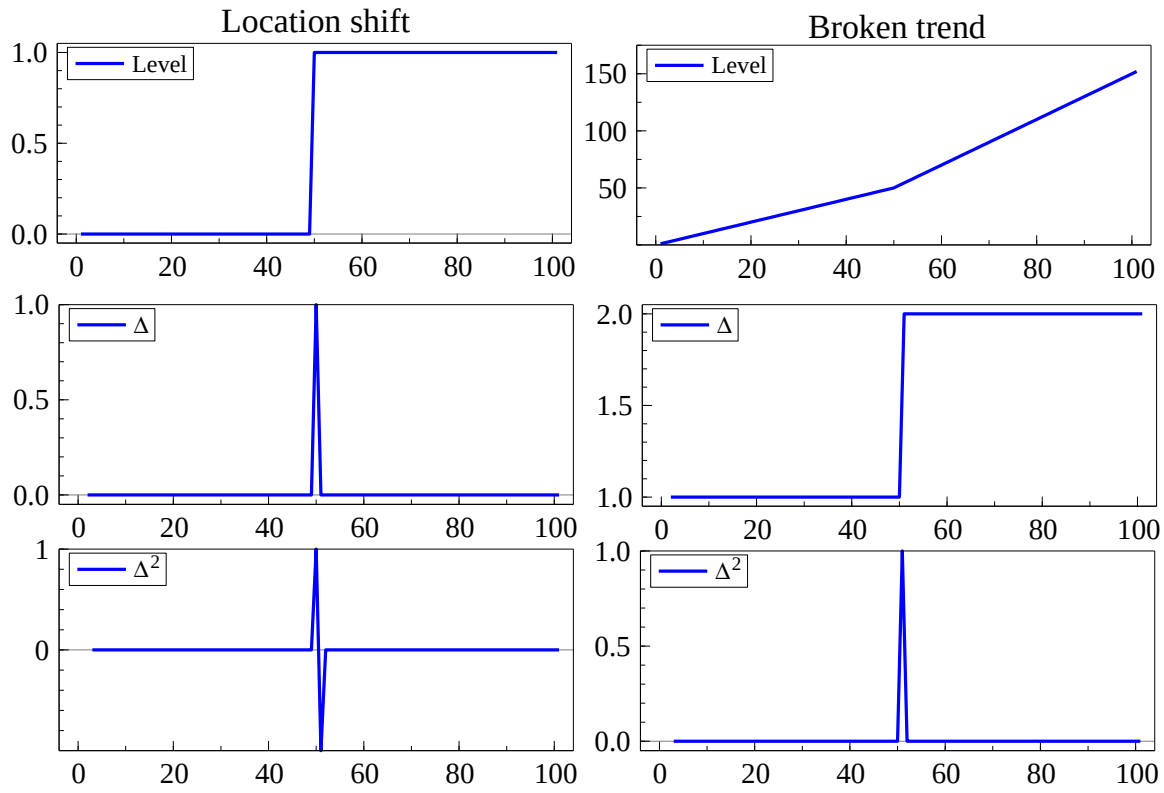

Figure 1 Location shifts and broken trends.

Also, most economic time series do not continuously accelerate – entailing a zero unconditional expectation of the second difference:

$$\mathsf{E}\left[\Delta^2 \mathbf{x_t}\right] = \mathbf{0}, \tag{29}$$

and suggesting the forecasting rule:

$$\Delta \widetilde{\mathbf{x}}_{T+1|T} = \Delta \mathbf{x_T}. \tag{30}$$

One key to the success of double differencing is that no deterministic terms remain, so that for time series like speculative prices, where no deterministic terms are present, 'random walk forecasts' will be equally hard to beat. However, as discussed below, differencing is incompatible with solutions to measurement errors as it exacerbates negative moving averages.

Nevertheless, there is a deeper reason why a forecast of the form (30) may generally perform well. Consider the in-sample DGP:

$$\Delta \mathbf{x_t} = \boldsymbol{\gamma_0} + \boldsymbol{\alpha_0}\left(\boldsymbol{\beta}_\mathbf{0}' \mathbf{x_{t-1}} - \boldsymbol{\mu_0}\right) + \boldsymbol{\Psi_0} \mathbf{z_t} + \boldsymbol{\epsilon_t}, \tag{31}$$

where $\boldsymbol{\epsilon_t} \sim \mathsf{IN_n}\left[\mathbf{0}, \boldsymbol{\Omega_\epsilon}\right]$ independently of all the included variables and their history, with population parameter values denoted by the subscript 0. Also, $\mathbf{z_t}$ denotes potentially many omitted I(0) effects, possibly all lagged (I(0), perhaps because of 'internal' cointegration, being differenced, or intrinsically stationary). The postulated econometric model is a VEqCM in $\mathbf{x_t}$:

$$\Delta \mathbf{x_T} = \boldsymbol{\gamma} + \boldsymbol{\alpha}\left(\boldsymbol{\beta}' \mathbf{x_{T-1}} - \boldsymbol{\mu}\right) + \mathbf{v_T},$$

and that model, estimated from T observations, is used for forecasting:

$$\Delta \widehat{\mathbf{x}}_{T+i|T+i-1} = \widehat{\boldsymbol{\gamma}} + \widehat{\boldsymbol{\alpha}}\left(\widehat{\boldsymbol{\beta}}' \mathbf{x_{T+i-1}} - \widehat{\boldsymbol{\mu}}\right). \tag{32}$$

Finally, over the forecast horizon, the DGP becomes:

$$\Delta \mathbf{x_{T+i}} = \boldsymbol{\gamma}_\mathbf{0}^* + \boldsymbol{\alpha}_\mathbf{0}^*\left((\boldsymbol{\beta}_\mathbf{0}^*)' \mathbf{x_{T+i-1}} - \boldsymbol{\mu}_\mathbf{0}^*\right) + \boldsymbol{\Psi}_\mathbf{0}^* \mathbf{z_{T+i}} + \boldsymbol{\epsilon_{T+i}}. \tag{33}$$

All the main sources of forecast error occur, given (33): stochastic and deterministic breaks, omitted variables, inconsistent parameter estimates, estimation uncertainty, and innovation errors: data measurement errors could be added. Thus, if $\Delta \mathbf{x}_{\mathbf{T+i}} - \boldsymbol{\Delta} \widehat{\mathbf{x}}_{\mathbf{T+i}|\mathbf{T+i-1}} = \mathbf{w}_{\mathbf{T+i}}$:

$$\mathbf{w}_{\mathbf{T+i}} = \boldsymbol{\gamma}_{\mathbf{0}}^{*} + \boldsymbol{\alpha}_{\mathbf{0}}^{*} \left((\boldsymbol{\beta}_{\mathbf{0}}^{*})' \mathbf{x}_{\mathbf{T+i-1}} - \boldsymbol{\mu}_{\mathbf{0}}^{*} \right) + \boldsymbol{\Psi}_{\mathbf{0}}^{*} \mathbf{z}_{\mathbf{T+i}} + \boldsymbol{\epsilon}_{\mathbf{T+i}} - \widehat{\boldsymbol{\gamma}} - \widehat{\boldsymbol{\alpha}} \left(\widehat{\boldsymbol{\beta}}' \mathbf{x}_{\mathbf{T+i-1}} - \widehat{\boldsymbol{\mu}} \right). \tag{34}$$

It is difficult to analyze (34) as its terms are not necessarily even $\mathsf{I}(0)$, but conditional on $(\mathbf{x}_{\mathbf{T+i-1}}, \mathbf{z}_{\mathbf{T+i-1}})$, $\mathbf{w}_{\mathbf{T+i}}$ has an approximate mean forecast error (using $\mathsf{E}[\widehat{\boldsymbol{\gamma}}] = \boldsymbol{\gamma}_{\mathbf{p}}$ etc.) of:

$$\begin{aligned} \mathsf{E}\left[\mathbf{w}_{\mathbf{T+i}} \mid \mathbf{x}_{\mathbf{T+i-1}}, \mathbf{z}_{\mathbf{T+i-1}} \right] &= \left(\boldsymbol{\gamma}_{\mathbf{0}}^{*} - \boldsymbol{\gamma}_{\mathbf{p}} \right) - \left(\boldsymbol{\alpha}_{\mathbf{0}}^{*} \boldsymbol{\mu}_{\mathbf{0}}^{*} - \boldsymbol{\alpha}_{\mathbf{p}} \boldsymbol{\mu}_{\mathbf{p}} \right) + \left[\boldsymbol{\alpha}_{\mathbf{0}}^{*} (\boldsymbol{\beta}_{\mathbf{0}}^{*})' - \boldsymbol{\alpha}_{\mathbf{p}} \boldsymbol{\beta}_{\mathbf{p}}' \right] \mathbf{x}_{\mathbf{T+i-1}} \\ &\quad + \boldsymbol{\Psi}_{\mathbf{0}}^{*} \mathsf{E}\left[\mathbf{z}_{\mathbf{T+i}} \mid \mathbf{x}_{\mathbf{T+i-1}}, \mathbf{z}_{\mathbf{T+i-1}} \right]. \end{aligned}$$

Also, neglecting parameter estimation uncertainty as $\mathsf{O}_p(T^{-1})$, $\mathbf{w}_{\mathbf{T+i}}$ has an approximate conditional error-variance matrix:

$$\mathsf{V}\left[\mathbf{w}_{\mathbf{T+i}} \mid \mathbf{x}_{\mathbf{T+i-1}}, \mathbf{z}_{\mathbf{T+i-1}} \right] = \boldsymbol{\Psi}_{\mathbf{0}}^{*} \mathsf{V}\left[\mathbf{z}_{\mathbf{T+i}} \mid \mathbf{x}_{\mathbf{T+i-1}}, \mathbf{z}_{\mathbf{T+i-1}} \right] \boldsymbol{\Psi}_{\mathbf{0}}^{*\prime} + \boldsymbol{\Omega}_{\boldsymbol{\epsilon}}, \tag{35}$$

and its conditional mean-square forecast error matrix is the sum of $\mathsf{E}[\mathbf{w}_{\mathbf{T+i}}|\mathbf{x}_{\mathbf{T+i-1}}, \mathbf{z}_{\mathbf{T+i-1}}]\mathsf{E}[\mathbf{w}_{\mathbf{T+i}}|\mathbf{x}_{\mathbf{T+i-1}}, \mathbf{z}_{\mathbf{T+i-1}}]'$ and (35).

Contrast using the sequence of $\Delta \mathbf{x}_{\mathbf{T+i-1}}$ to forecast $\Delta \mathbf{x}_{\mathbf{T+i}}$, as in (30):

$$\Delta \widetilde{\mathbf{x}}_{T+i|T+i-1} = \Delta \mathbf{x}_{\mathbf{T+i-1}}. \tag{36}$$

Because of (33), $\Delta \mathbf{x}_{\mathbf{T+i-1}}$ is in fact (for $i > 1$):

$$\Delta \mathbf{x}_{\mathbf{T+i-1}} = \boldsymbol{\gamma}_{\mathbf{0}}^{*} + \boldsymbol{\alpha}_{\mathbf{0}}^{*} \left((\boldsymbol{\beta}_{\mathbf{0}}^{*})' \mathbf{x}_{\mathbf{T+i-2}} - \boldsymbol{\mu}_{\mathbf{0}}^{*} \right) + \boldsymbol{\Psi}_{\mathbf{0}}^{*} \mathbf{z}_{\mathbf{T+i-1}} + \boldsymbol{\epsilon}_{\mathbf{T+i-1}}. \tag{37}$$

Thus, (37) shows that, without the economists needing to know the causal variables or the structure of the economy, $\Delta \mathbf{x}_{\mathbf{T+i-1}}$ **reflects all the effects in the DGP**, including all parameter changes, with no omitted variables and no estimation required at all. However, there are two drawbacks: the unwanted presence of $\boldsymbol{\epsilon}_{T+i-1}$ in (37), which doubles the innovation error variance; and all variables are lagged one extra period, which adds the 'noise' of many $\mathsf{I}(-1)$ effects. Thus, there is a clear trade-off between using the carefully modelled (32) and the 'naive' predictor (36). In forecasting competitions across many states of nature with structural breaks and complicated DGPs, it is easy to see why $\Delta \mathbf{x}_{\mathbf{T+i-1}}$ may win.

Let $\Delta \mathbf{x}_{\mathbf{T+i}} - \boldsymbol{\Delta} \widetilde{\mathbf{x}}_{\mathbf{T+i}|\mathbf{T}} = \mathbf{u}_{\mathbf{T+i}}$, then:

$$\begin{aligned} \mathbf{u}_{\mathbf{T+i}} &= \boldsymbol{\gamma}_{\mathbf{0}}^{*} + \boldsymbol{\alpha}_{\mathbf{0}}^{*} \left((\boldsymbol{\beta}_{\mathbf{0}}^{*})' \mathbf{x}_{\mathbf{T+i-1}} - \boldsymbol{\mu}_{\mathbf{0}}^{*} \right) + \boldsymbol{\Psi}_{\mathbf{0}}^{*} \mathbf{z}_{\mathbf{T+i-1}} + \boldsymbol{\epsilon}_{\mathbf{T+i}} \\ &\quad - \left[\boldsymbol{\gamma}_{\mathbf{0}}^{*} + \boldsymbol{\alpha}_{\mathbf{0}}^{*} \left((\boldsymbol{\beta}_{\mathbf{0}}^{*})' \mathbf{x}_{\mathbf{T+i-2}} - \boldsymbol{\mu}_{\mathbf{0}}^{*} \right) + \boldsymbol{\Psi}_{\mathbf{0}}^{*} \mathbf{z}_{\mathbf{T+i-1}} + \boldsymbol{\epsilon}_{\mathbf{T+i-1}} \right] \\ &= \boldsymbol{\alpha}_{\mathbf{0}}^{*} (\boldsymbol{\beta}_{\mathbf{0}}^{*})' \boldsymbol{\Delta} \mathbf{x}_{\mathbf{T+i-1}} + \boldsymbol{\Psi}_{\mathbf{0}}^{*} \boldsymbol{\Delta} \mathbf{z}_{\mathbf{T+i}} + \boldsymbol{\Delta} \boldsymbol{\epsilon}_{\mathbf{T+i}}. \end{aligned} \tag{38}$$

All terms in the last line must be $\mathsf{I}(-1)$, so will be very 'noisy', but no systematic failure should result. Indeed:

$$\mathsf{E}\left[\mathbf{u}_{\mathbf{T+i}} \right] = \boldsymbol{\alpha}_{\mathbf{0}}^{*} \mathsf{E}\left[(\boldsymbol{\beta}_{\mathbf{0}}^{*})' \boldsymbol{\Delta} \mathbf{x}_{\mathbf{T+i-1}} \right] + \boldsymbol{\Psi}_{\mathbf{0}}^{*} \mathsf{E}\left[\Delta \mathbf{z}_{\mathbf{T+i}} \right] + \mathsf{E}\left[\Delta \boldsymbol{\epsilon}_{\mathbf{T+i}} \right] = \boldsymbol{\alpha}_{\mathbf{0}}^{*} (\boldsymbol{\beta}_{\mathbf{0}}^{*})' \boldsymbol{\gamma}_{\mathbf{0}}^{*\prime} = \mathbf{0}.$$

Neglecting covariances, we have:

$$\begin{aligned} \mathsf{V}\left[\mathbf{u}_{\mathbf{T+i}} \right] &= \mathsf{V}\left[\boldsymbol{\alpha}_{\mathbf{0}}^{*} (\boldsymbol{\beta}_{\mathbf{0}}^{*})' \boldsymbol{\Delta} \mathbf{x}_{\mathbf{T+i-1}} \right] + \mathsf{V}\left[\boldsymbol{\Psi}_{\mathbf{0}}^{*} \boldsymbol{\Delta} \mathbf{z}_{\mathbf{T+i}} \right] + \mathsf{V}\left[\Delta \boldsymbol{\epsilon}_{\mathbf{T+i}} \right] \\ &= \boldsymbol{\alpha}_{\mathbf{0}}^{*} (\boldsymbol{\beta}_{\mathbf{0}}^{*})' \mathsf{V}\left[\Delta \mathbf{x}_{\mathbf{T+i-1}} \right] \boldsymbol{\beta}_{\mathbf{0}}^{*} \boldsymbol{\alpha}_{\mathbf{0}}^{*\prime} + \boldsymbol{\Psi}_{\mathbf{0}}^{*} \mathsf{V}\left[\Delta \mathbf{z}_{\mathbf{T+i}} \right] \boldsymbol{\Psi}_{\mathbf{0}}^{*\prime} + \mathbf{2} \boldsymbol{\Omega}_{\boldsymbol{\epsilon}} \end{aligned} \tag{39}$$

which is the mean-square error matrix because $\mathsf{E}[\mathbf{u}_{\mathbf{T+i}}] = \mathbf{0}$. Conventional analysis notes the doubling of $\mathbf{\Omega}_{\boldsymbol{\epsilon}}$ in (39) relative to (35). However, when $\{\mathbf{z_t}\}$ is a stationary vector autoregression (say):

$$\mathbf{z_t} = \mathbf{\Gamma z_{t-1}} + \boldsymbol{\eta}_\mathbf{t} \text{ where } \boldsymbol{\eta}_\mathbf{t} \sim \mathsf{IN_k}[\mathbf{0}, \mathbf{\Omega}_{\boldsymbol{\eta}}],$$

then:

$$\mathsf{V}[\mathbf{z_t}] = \mathbf{\Gamma}\mathsf{V}[\mathbf{z_t}]\mathbf{\Gamma}' + \mathbf{\Omega}_{\boldsymbol{\eta}},$$

and:

$$\mathsf{V}[\Delta\mathbf{z_t}] = (\mathbf{\Gamma} - \mathbf{I_k})\mathsf{V}[\mathbf{z_t}](\mathbf{\Gamma} - \mathbf{I_k})' + \mathbf{\Omega}_{\boldsymbol{\eta}}$$

so that:

$$\begin{aligned}\mathsf{V}[\Delta\mathbf{z}_{\mathbf{T+i}}] - \mathsf{V}[\mathbf{z}_{\mathbf{T+i}}] &= (\mathbf{\Gamma} - \mathbf{I_k})\mathsf{V}[\mathbf{z}_{\mathbf{T+i}}](\mathbf{\Gamma} - \mathbf{I_k})' - \mathbf{\Gamma}\mathsf{V}[\mathbf{z}_{\mathbf{T+i}}]\mathbf{\Gamma}' \\ &= \mathsf{V}[\mathbf{z}_{\mathbf{T+i}}] - \mathbf{\Gamma}\mathsf{V}[\mathbf{z}_{\mathbf{T+i}}] - \mathsf{V}[\mathbf{z}_{\mathbf{T+i}}]\mathbf{\Gamma}'\end{aligned}$$

which could attain a maximum of $\mathsf{V}[\mathbf{z}_{\mathbf{T+i}}]$ when $\{\mathbf{z_t}\}$ is white noise ($\mathbf{\Gamma} = \mathbf{0}$), or approach $-\mathsf{V}[\mathbf{z}_{\mathbf{T+i}}]$ when $\{\mathbf{z_t}\}$ is highly autoregressive ($\mathbf{\Gamma} \simeq \mathbf{I_k}$). Thus, the overall error variance in (39) will not necessarily double relative to (35), and could be smaller in sufficiently badly specified VEqCMs.

5.2 Rapid updating

An alternative to over-differencing is more rapid updating of the coefficients of the deterministic terms, possibly using different estimators for forecasting. Thus, in a 'non-causal' representation, consider a short moving average of past actual growth rates, so:

$$\widetilde{\Delta\mathbf{x}}_{T+1|T} = \widetilde{\boldsymbol{\tau}}_T \tag{40}$$

where:

$$\widetilde{\boldsymbol{\tau}}_T = \frac{1}{m+1}\sum_{i=0}^{m}\Delta\mathbf{x}_{\mathbf{T-i}}. \tag{41}$$

Then:

$$(m+1)\widetilde{\boldsymbol{\tau}}_T = \sum_{i=0}^{m}\Delta\mathbf{x}_{\mathbf{T-i}} = \mathbf{\Delta x_T} - \mathbf{\Delta x}_{\mathbf{T-(m+1)}} + (\mathbf{m+1})\widetilde{\boldsymbol{\tau}}_{\mathbf{T-1}},$$

so:

$$\widetilde{\boldsymbol{\tau}}_T = \widetilde{\boldsymbol{\tau}}_{T-1} + \frac{1}{(m+1)}\Delta\Delta_{(m+1)}\mathbf{x_T},$$

reflecting aspects of Kalman filtering. When $m = 0$:

$$\widetilde{\Delta\mathbf{x}}_{T+1|T} = \Delta\mathbf{x_T},$$

which reproduces the DDV as corresponding to updating the intercept by the latest 'surprise', $\Delta^2\mathbf{x_T}$. Larger values of m will 'smooth' intercept estimates, but adapt more slowly: using $m = T-1$ essentially delivers the OLS estimates, which do not adapt. For forecasting from quarterly data using $m = 3$:

$$\widetilde{\boldsymbol{\tau}}_T = \frac{1}{4}(\Delta\mathbf{x_T} + \cdots + \mathbf{\Delta x_{T-3}}) = \frac{1}{4}\Delta_4\mathbf{x_T},$$

which is the previous average annual growth. So long as breaks are not too frequent, and the variables to be forecast do not accelerate, such devices seem likely to work reasonably well in avoiding systematic forecast failure.

Implicit in (40):

$$\widetilde{\boldsymbol{\tau}}_T = \widetilde{\boldsymbol{\gamma}}_T - \boldsymbol{\alpha}\widetilde{\boldsymbol{\mu}}_{\mathbf{T}},$$

and so it reflects changes in either source of intercept shift. Once a more causally-based model is used, that mapping ceases, so implementing an analogous notion requires care. The basic problem is that if such corrections work well when a model is mis-specified, they cannot be appropriate when it is valid for the same observed change in growth: the latter case mis-attributes changed observed growth to a shift in $\boldsymbol{\tau}$ whereas it will be captured by other regressors. A lack of orthogonality between the various 'explanatory components' is the source of this difficulty: changes in one variable are confounded with resulting changes in the growth rates of others (see Bewley, 2000, for an alternative parameterization that seeks ro resolve this problem).

5.3 Forecast-error based adaptation

Consequently, only forecast-error based information, which reflects the problems of the model, not the changes in the data, can be used to correct breaks in econometric systems. Apart from ICs (which add back recent errors, and are also susceptible to smoothing), one of the most famous 'forecast-error correction' mechanisms (FErCMs) is the exponentially weighted moving average (EWMA), so we consider its possible transmogrification to econometric systems.

The EWMA recursive updating formula is, for $\lambda \in (0,1)$ and a scalar time series $\{y_t\}$:

$$\widehat{y}_{T+h|T} = (1-\lambda)\sum_{j=0}^{\infty}\lambda^j y_{T-j},$$

so (e.g.):

$$\widehat{y}_{T+1|T} = (1-\lambda)\,y_T + \lambda\widehat{y}_{T|T-1} = y_T - \lambda\left(y_T - \widehat{y}_{T|T-1}\right), \tag{42}$$

with start-up value $\widehat{y}_1 = y_1$. Hence, for an origin T, $\widehat{y}_{T+h|T} = \widehat{y}_{T+1|T}$ for all h. One can view this method as 'correcting' a random-walk forecast by the latest forecast error $(y_T - \widehat{y}_{T|T-1})$:

$$\widehat{\Delta y}_{T+1|T} = -\lambda\left(y_T - \widehat{y}_{T|T-1}\right), \tag{43}$$

possibly seen as approximating the ARIMA(0,1,1):

$$\Delta y_t = \varepsilon_t - \theta\varepsilon_{t-1}, \tag{44}$$

so the second term in (42) seeks to offset that in (44):

$$\Delta y_{T+1} - \widehat{\Delta y}_{T+1|T} = \varepsilon_{T+1} - \theta\varepsilon_T + \lambda\left(y_T - \widehat{y}_{T|T-1}\right).$$

Consequently, (42) could be seen as being designed for data measured with error, where the underlying model was $\Delta y_t^* = v_t$ with $y_t = y_t^* + w_t$ so that:

$$\Delta y_t = \Delta y_t^* + \Delta w_t = (v_t + w_t) - w_{t-1}.$$

Any shift in the mean of $\{y\}$ will eventually feed through to the forecasts from (42): adding back a damped function of recent forecast errors ought, therefore, to be productive when location shifts are common. The speed with which adjustment occurs depends on the degree of damping, λ, where $\lambda = 0$ corresponds to a random walk forecast. The choice of a large λ prevents the predictor extrapolating the 'noise' in the latest observation, but when there is a shift in mean, the closer λ is to zero the more quickly a break will be assimilated in the forecasts.

5.3.1 The relation of EWMA and IC

Four components seem to contribute to the forecasting success of EWMA:

- adapting the next forecast by the previous forecast error;
- differencing to adjust to location shifts;
- the absence of deterministic terms which could go awry;
- rapidly adaptive when λ is small.

The correction of a forecast by a previous forecast error is reminiscient of intercept correction. However, EWMA differs from IC by the sign and size of the damping factor, $-\lambda$ in place of unity, so may not face the latter's problems when there are large measurement errors at the forecast origin. To investigate the implications of this sign change, consider a vector generalization of (43) using the forecast from (45), (abstracting from parameter estimation):

$$\widehat{\Delta \mathbf{x}}_{T+1|T} = \boldsymbol{\gamma} + \left(\boldsymbol{\alpha}\boldsymbol{\beta}'\mathbf{x_T} - \boldsymbol{\alpha}\boldsymbol{\mu}\right), \tag{45}$$

when augmented by the forecast-error correction:

$$\overline{\Delta \mathbf{x}}_{T+1|T} = \widehat{\Delta \mathbf{x}}_{T+1|T} - \boldsymbol{\Lambda}\left(\mathbf{x_T} - \overline{\mathbf{x}}_{\mathbf{T}|\mathbf{T-1}}\right). \tag{46}$$

Assuming the VEqCM (24) was congruent in-sample, then using:

$$\overline{\mathbf{x}}_{T|T-1} = \overline{\Delta \mathbf{x}}_{T|T-1} + \mathbf{x_{T-1}} = \mathbf{x_{T-1}} + \boldsymbol{\gamma} + \left(\boldsymbol{\alpha}\boldsymbol{\beta}'\mathbf{x_{T-1}} - \boldsymbol{\alpha}\boldsymbol{\mu}\right),$$

leads to:

$$\mathbf{x_T} - \overline{\mathbf{x}}_{\mathbf{T}|\mathbf{T-1}} = \boldsymbol{\Delta}\mathbf{x_T} - \boldsymbol{\gamma} - \boldsymbol{\alpha}\left(\boldsymbol{\beta}'\mathbf{x_{T-1}} - \boldsymbol{\mu}\right) = \boldsymbol{\Delta}\mathbf{x_T} - \widehat{\boldsymbol{\Delta}\mathbf{x}}_{\mathbf{T}|\mathbf{T-1}},$$

which is the last in-sample 1-step residual, $\widehat{\boldsymbol{\epsilon}}_T$. Thus, letting $\widehat{\boldsymbol{\epsilon}}_{T+1} = \Delta\mathbf{x_{T+1}} - \widehat{\boldsymbol{\Delta}\mathbf{x}}_{\mathbf{T+1}|\mathbf{T}}$:

$$\Delta\mathbf{x_{T+1}} - \overline{\boldsymbol{\Delta}\mathbf{x}}_{\mathbf{T+1}|\mathbf{T}} = \widehat{\boldsymbol{\epsilon}}_{\mathbf{T+1}} + \boldsymbol{\Lambda}\widehat{\boldsymbol{\epsilon}}_{\mathbf{T}},$$

so $\boldsymbol{\Lambda} = -\mathbf{I_n}$ corresponds to the IC for 'setting the forecasts back on track' at the forecast origin. The sign change is not due to IC being an autoregressive, rather than a moving-average, correction: rather, the aim of the IC is to offset a location shift, whereas EWMA seeks to offset a previous measurement error, using differencing to remove location shifts. Thus, we see an important *caveat* to the explanations fo the empirical success of ICs discussed in Clements and Hendry (1999, Ch.6): some of the potential roles conflict. In particular, to offset previous mis-specifications or measurement errors requires the opposite sign to that for offsetting breaks.

5.3.2 Adapting EWMA for growth changes

The absence of any deterministic terms in (43) entails that if the data are growing, systematic under-prediction may occur. This last difficulty could be circumvented by an extra degree of differencing as in the type of model discussed by Harvey and Shephard (1992) (so y_t in (42) becomes the growth rate), or alternatively by letting:

$$\widetilde{\Delta y}_{T+1|T} = \widetilde{\gamma}_T - \lambda\left(y_T - \widetilde{y}_{T|T-1}\right) \tag{47}$$

where:

$$\widetilde{\gamma}_T = \widetilde{\gamma}_{T-1} + \frac{1}{(m+1)}\Delta_{(m+1)}\Delta y_T. \tag{48}$$

Notice that $\widetilde{\gamma}_{T-1}$ could be based on all the in-sample data, switching to (48) only when forecasting. However, $m = 0$ (say) enforces complete adaptation to the latest 'surprise' $\Delta^2 y_T$, which could be noisy. The 'combined' device in (47) both corrects recent past errors and adjusts rapidly to changes in observed growth irrespective of whether that corresponds to changes in $\boldsymbol{\gamma}$ in the DGP, or is an induced effect from shifts in $\boldsymbol{\mu}$.

Vector generalizations of (47) and (48) are straight-forward—the former becomes:

$$\widetilde{\Delta \mathbf{x}}_{T+1|T} = \widetilde{\boldsymbol{\gamma}}_T - \boldsymbol{\Lambda} \left(\mathbf{x_T} - \widetilde{\mathbf{x}}_{\mathbf{T}|\mathbf{T-1}} \right) \tag{49}$$

where $\boldsymbol{\Lambda}$ could be diagonal, denoted ADV for adaptive DVAR. Then, (49) generalizes the simplest DVAR-based forecast:

$$\widetilde{\Delta \mathbf{x}}_{T+1|T} = \boldsymbol{\gamma}, \tag{50}$$

and is also similar to (22) when that equation is written as (for known in-sample parameters):

$$\widehat{\Delta \mathbf{x}}_{T+1|T} = \boldsymbol{\gamma} + \left(\boldsymbol{\alpha} \boldsymbol{\beta}' \mathbf{x_T} - \boldsymbol{\alpha} \boldsymbol{\mu} \right), \tag{51}$$

but with the equilibrium correction in (22) replaced by forecast-error correction in (49). Alternatively, combination leads to (46) above, which augments (51) by the last term in (49).

6 Empirical illustration of UK M1

The two 'forecasting' models of UK M1 in Hendry and Mizon (1993) and Hendry and Doornik (1994) respectively illustrate several of these phenomena (see those papers for details of the models: other closely related studies include Hendry, 1979, Hendry and Ericsson, 1991, Boswijk, 1992, Johansen, 1992, Paruolo, 1996 and Rahbek, Kongsted and Jørgensen, 1999). The data are quarterly, seasonally-adjusted, time series over 1963(1)–1989(2), defined as:

M	nominal M1,
I	real total final expenditure (*TFE*) at 1985 prices,
P	the *TFE* deflator,
R_{la}	the three-month local authority interest rate,
R_o	learning-adjusted own interest rate,
R_{net}	$R_{la} - R_o$.

The first model is based on using the competitive interest rate R_{la}, and the second on the opportunity-cost measure R_{net} appropriate after the Banking Act of 1984 legalized interest payments on chequing accounts. To simplify the results, we first consider only the money-demand model, then turn briefly to system behaviour. In both cases, 'forecasts' are over the five years 1984(3)–1989(2), or subsets thereof.

6.1 Single-equation results

The first step is to illustrate that the Banking Act corresponded to an equilibrium-mean shift relative to the model based on R_{la}. The own rate, R_o has a mean of approximately 0.075 over the forecast horizon, and a shift indicator $1_{\{t>1985(2)\}}$ times that mean closely approximates the actual time path of that variable: see figure 2, panel a. Thus, subtracting $0.075 \times 1_{\{t>1985(2)\}}$ from R_{la} is close to R_{net} (denoted R^c_{la} in figure 2b): it is clear why an intercept correction should perform well after 1985(4). Next, over the forecast horizon, the moving average growth rate of real money shifted dramatically relative to the recursively estimated historical mean growth rate (see figure 2c): this reflects both effects

in $\nabla\boldsymbol{\gamma}^* - \boldsymbol{\alpha}\boldsymbol{\nabla}\boldsymbol{\mu}^*$, even if $\nabla\boldsymbol{\gamma}^* = \mathbf{0}$ so the 'fundamental' growth rate is unchanged. Since that observed growth mimics the 'missing ingredient' in a univariate forecasting device, the second adaptation above should be successful in that context. Finally, the estimate of the original equilibrium mean, $\widehat{\boldsymbol{\mu}}$ based on R_{la} is quite sensible (see figure 2d), and shows no signs of a shift, whereas $\widetilde{\boldsymbol{\mu}}$ based on R_{net}, does shift. At first sight, that may seem counterintuitive, but it occurs precisely because the opportunity costs have shifted dramatically, so $\widehat{\boldsymbol{\mu}}$ does not reflect that shift, thereby causing forecast failure. Consequently, real money and R_{net} must co-break, as illustrated in Clements and Hendry (1999, Ch. 9).

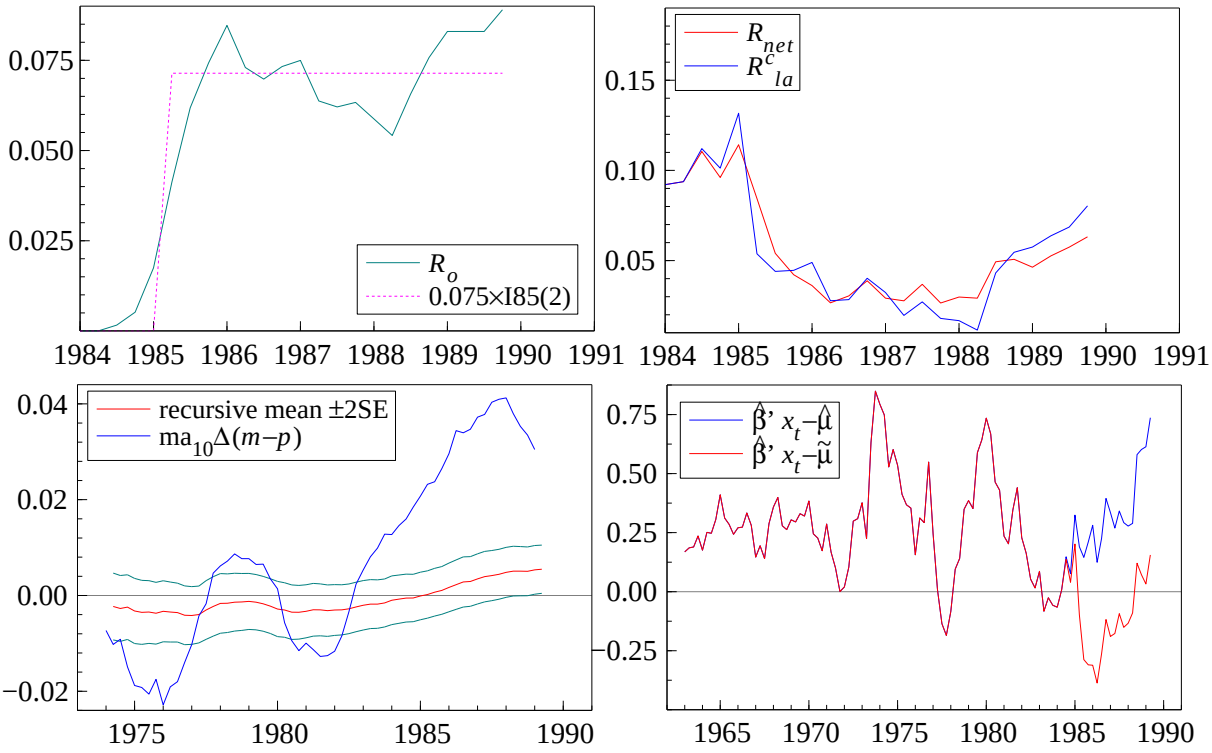

Figure 2 Effects of the 1984 Banking Act on UK M1 .

Figure 3a shows the dismal performance on 20 1-step forecasts of the Hendry and Mizon (1993) model: almost none of the $\pm 2\widehat{\sigma}$ error bars includes the outcome, and a large fall is forecast during the largest rise experienced historically, so the level is dramatically underestimated.

For comparison, the 20 1-step forecasts from the first difference of that original model are shown in figure 3b: there is a very substantial improvement, with no systematic under-forecasting, suggesting that the first proposed adaptation can be effective in the face of equilibrium-mean shifts (all the panels are on the same scale, so the corresponding increase in the interval forecasts is also clear).

Next, correcting the original model (i.e., with $\widehat{\boldsymbol{\mu}}$ based on R_{la}) by an estimate of the changed transient growth rate, namely (schematically):

$$\overline{\Delta \mathbf{x}}_{T+1|T} - \widetilde{\boldsymbol{\gamma}}_T = \widehat{\boldsymbol{\alpha}} \left(\widehat{\boldsymbol{\beta}}' \mathbf{x}_{\mathbf{T}} - \widehat{\boldsymbol{\mu}} \right) + \cdots ,$$

using:

$$\widetilde{\boldsymbol{\gamma}}_T = \frac{1}{4} \sum_{i=0}^{3} \Delta \mathbf{x}_{\mathbf{T-i}}, \tag{52}$$

is also effective, as shown in figure 3c, although it can be seen to be drifting off course at the end once economic agents have adjusted to their new environment, and the observed growth rate reverts to $\boldsymbol{\gamma}$ (which no longer reflects $\boldsymbol{\alpha}\boldsymbol{\nabla}\boldsymbol{\mu}^*$).

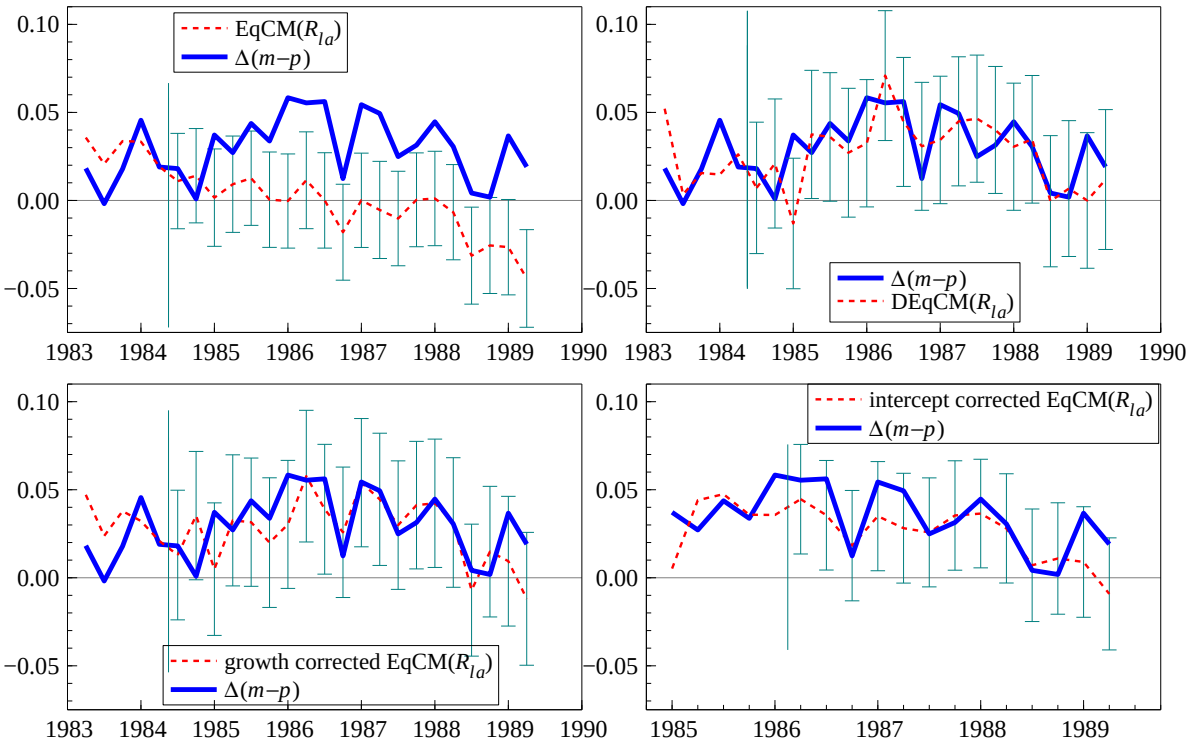

Figure 3 Forecasts of UK M1 adapting the R_{la}-based model.

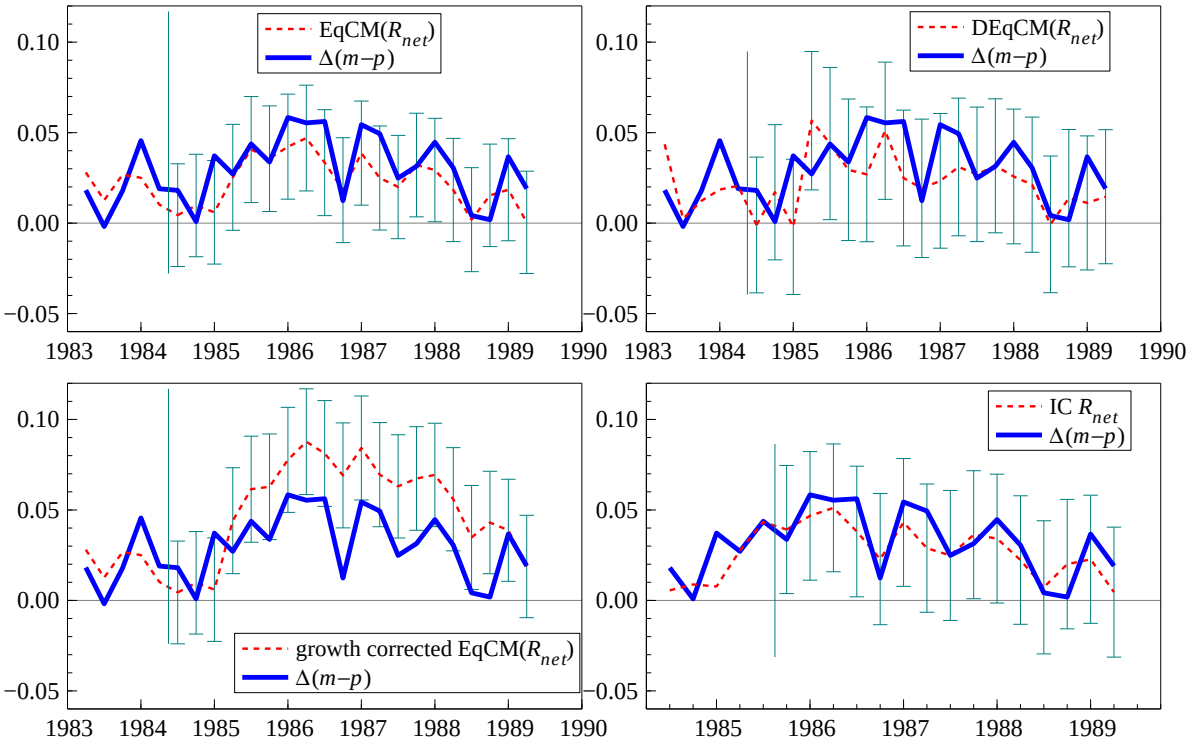

Figure 4 Forecasts of UK M1 adapting the R_{net}-based model.

Finally, figure 3d records the IC adjusted forecasts over the shorter horizon from 1986(1) (so the adjustment can be estimated): even so, it is the least successful of these three adaptations.

Figure 4a shows the good performance on 20 1-step forecasts of the 'correct' model (i.e., based on R_{net}). Since one cannot know whether a given model is robust to a break, the effects of the three

adaptations applied to the R_{net} model are also worth investigating. Differencing the EqCM produces similar forecasts to the EqCM itself as shown in figure 4b, but with larger error bars; however, even for a 'correct specification', the costs of that strategy do not seem to be too high. The same cannot be said for the results obtained by correcting using $\widetilde{\gamma}_T$ in figure 4c, which confirms the anticipated poor performance: the regressors already fully account for the increased growth, so that strategy is likely to be useful only for univariate models. Finally, an IC is insignificant if added to the model using R_{net} and so has little impact on the forecasts beyond an increase in the error bars (see figure 4).

For comparison, forecasts based on the most naive device, the DDV, are shown in figure 5 panel a. The DDV actually has a smaller mean error than the 'correct' model (-0.01% as against 0.9%), but a much larger standard deviation (2.25% against 1.19%), so the benefits of causal information are marked.[4] The ADV forecasts (based on (49) with $\Lambda = 0$) using $\widetilde{\gamma}_T$ from (52), and shown in figure 5c, are distinctly better than the DDV (RMSFE of 1.8% as against 2.25%). This is also true of the ADV and DDV forecasts for R_{net} shown in figure 5 panels b and d (RMSFE of 1.5% as against 1.9%). Thus, while double differencing is highly adaptive when a break occurs, the additional error variance at all points seems to more than offset its advantage in comparison to the smoother adaptation used here.

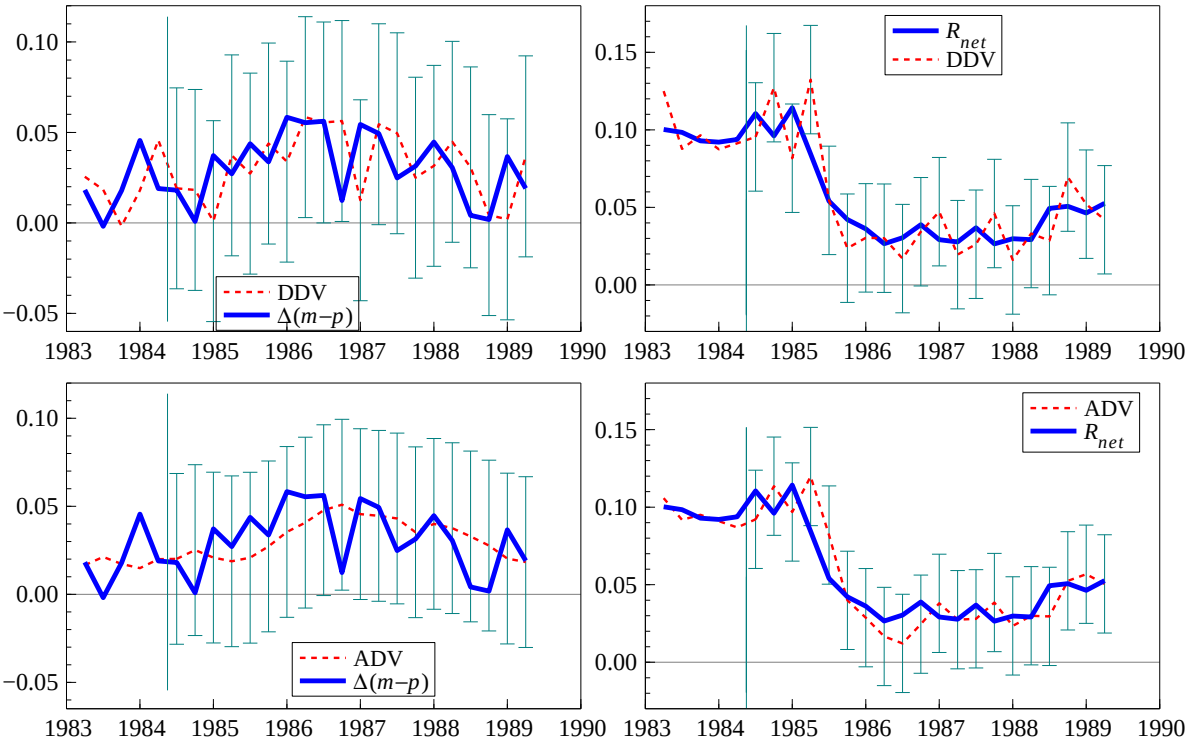

Figure 5 DDV and ADV forecasts of UK M1.

6.2 System behaviour

The above are all essentially single-equation forecasts, although the DDV and ADV devices are unaltered by being embedded in a system. In a system context, however, the break in the money-demand equation in the first VEqCM based on R_{la} becomes, in the second VEqCM, a shift in the R_{net} equation—which in turn could not be forecast accurately, as can be seen in figure 6, panels a and b (the outcomes for *TFE* and Δp are omitted).

[4] Subject to the *caveats* that the former uses current-dated variables in its 'forecasts', and the error bars on the DDV graph fail to correct for the negative residual serial correlation.

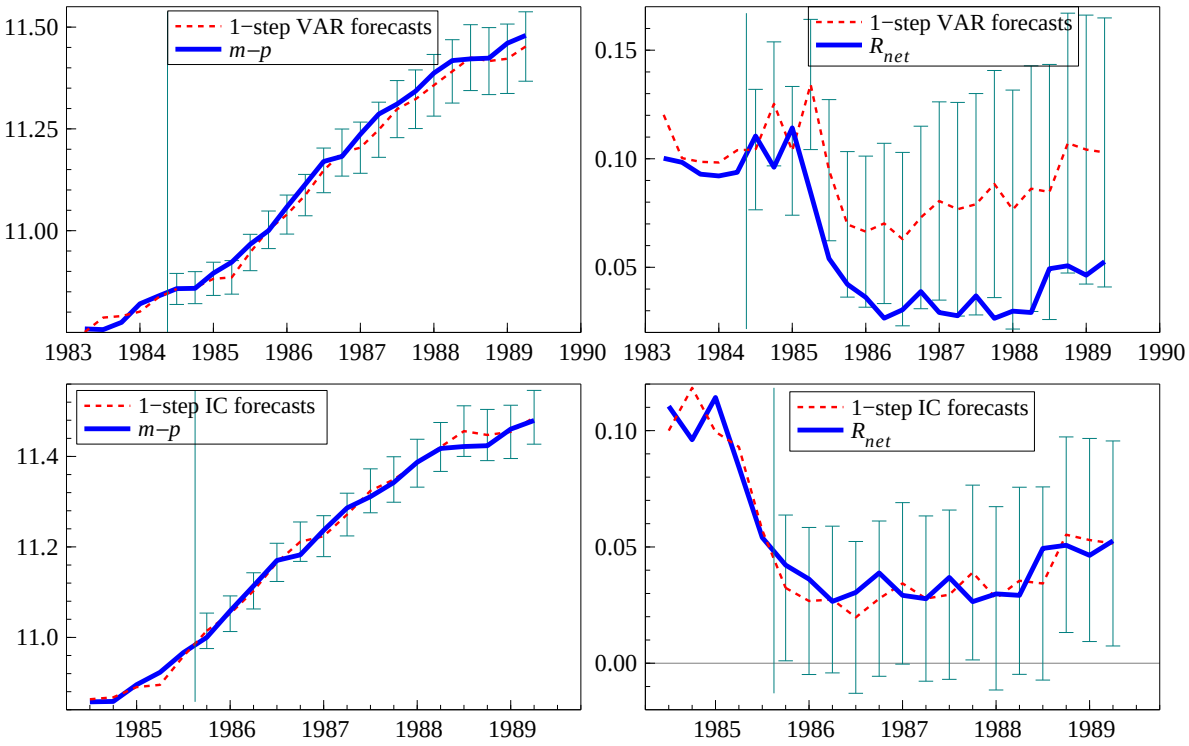

Figure 6 System forecasts from two 4-variable VARs of UK M1.

Nevertheless, the adaptations generalize to the other equations of theses systems, and have corresponding impacts, illustrated in figure 5 for R_{net}. As another example, when co-breaking is known, so R_{net} is the only equation for which an IC is required, the outcomes in figure 6, panels c and d, result: R_{net} is accurately forecast, with perceptible improvements in the interval forecasts for real money (and *TFE*, though not shown). However, the ADV for R_{net} achieves a similarly outcome (RMSFE of 0.8% for the IC as against 1.0% for the ADV over 1985(4)–1989(2)), but applicable over a longer forecast horizon.

Forecasting volatility

Reconsider a GARCH(1,1) process where $\varphi_1 + \varphi_2 < 1$:

$$\sigma_t^2 = \varphi_0 + \varphi_1 u_{t-1}^2 + \varphi_2 \sigma_{t-1}^2. \tag{53}$$

The long-run variance is $\omega = \varphi_0/(1-\varphi_1-\varphi_2) > 0$ which implies that (53) is an equilibrium-correction model, and hence is not robust to shifts in ω, but may be resilient to shifts in φ_1 or φ_2 which leave ω unaltered, as those only impact on 'mean zero' terms:

$$\sigma_t^2 = \omega + \varphi_1 \left(u_{t-1}^2 - \sigma_{t-1}^2\right) + (\varphi_1 + \varphi_2) \left(\sigma_{t-1}^2 - \omega\right).$$

A forecast of next period's volatility would use:

$$\widehat{\sigma}_{T+1|T}^2 = \widehat{\omega} + \widehat{\varphi}_1 \left(\widehat{u}_T^2 - \widehat{\sigma}_T^2\right) + (\widehat{\varphi}_1 + \widehat{\varphi}_2) \left(\widehat{\sigma}_T^2 - \widehat{\omega}\right). \tag{54}$$

Then (54) confronts every problem noted above for forecasts of means: potential breaks in ω, φ_1, φ_2, mis-specification of the variance evolution (perhaps a different functional form), estimation uncertainty, etc.

The 1-step ahead forecast-error taxonomy takes the following form after a shift in φ_0, φ_1, φ_2 to φ_0^*, φ_1^*, φ_2^* at T to:

$$\sigma_{T+1}^2 = \omega^* + \varphi_1^* \left(u_T^2 - \sigma_T^2\right) + \left(\varphi_1^* + \varphi_2^*\right) \left(\sigma_T^2 - \omega^*\right),$$

so that letting the subscript $_p$ denote the plim:

$$\begin{aligned}
\sigma_{T+1}^2 - \widehat{\sigma}_{T+1|T}^2 \;=\; & \left(1 - \left(\varphi_1^* + \varphi_2^*\right)\right)\left(\omega^* - \omega\right) && \text{long-run mean shift, [1]}\\
& + \left(1 - \left(\widehat{\varphi}_1 + \widehat{\varphi}_2\right)\right)\left(\omega - \omega_p\right) && \text{long-run mean inconsistency, [2]}\\
& + \left(1 - \left(\widehat{\varphi}_1 + \widehat{\varphi}_2\right)\right)\left(\omega_p - \widehat{\omega}\right) && \text{long-run mean variability, [3]}\\
& + \left(\varphi_1^* - \varphi_1\right)\left(u_T^2 - \sigma_T^2\right) && \varphi_1 \text{ shift, [4]}\\
& + \left(\varphi_1 - \varphi_{1,p}\right)\left(u_T^2 - \sigma_T^2\right) && \varphi_1 \text{ inconsistency, [5]}\\
& + \left(\varphi_{1,p} - \widehat{\varphi}_1\right)\left(u_T^2 - \sigma_T^2\right) && \varphi_1 \text{ variability, [6]}\\
& + \widehat{\varphi}_1 \left(u_T^2 - \mathsf{E_T}\left[\widehat{\mathsf{u}}_{\mathsf{T}}^2\right]\right) && \text{impact inconsistency, [7]}\\
& + \widehat{\varphi}_1 \left(\mathsf{E_T}\left[\widehat{\mathsf{u}}_{\mathsf{T}}^2\right] - \widehat{\mathsf{u}}_{\mathsf{T}}^2\right) && \text{impact variability, [8]}\\
& + \left[\left(\varphi_1^* + \varphi_2^*\right) - \left(\varphi_1 + \varphi_2\right)\right]\left(\sigma_T^2 - \omega\right) && \text{variance shift, [9]}\\
& + \left[\left(\varphi_1 + \varphi_2\right) - \left(\varphi_{1,p} + \varphi_{2,p}\right)\right]\left(\sigma_T^2 - \omega\right) && \text{variance inconsistency, [10]}\\
& + \left[\left(\varphi_{1,p} + \varphi_{2,p}\right) - \left(\widehat{\varphi}_1 + \widehat{\varphi}_2\right)\right]\left(\sigma_T^2 - \omega\right) && \text{variance variability, [11]}\\
& + \widehat{\varphi}_2 \left(\sigma_T^2 - \mathsf{E_T}\left[\widehat{\sigma}_{\mathsf{T}}^2\right]\right) && \sigma_T^2 \text{ inconsistency, [12]}\\
& + \widehat{\varphi}_2 \left(\mathsf{E_T}\left[\widehat{\sigma}_{\mathsf{T}}^2\right] - \widehat{\sigma}_{\mathsf{T}}^2\right) && \sigma_T^2 \text{ variability, [13].}
\end{aligned}$$

The first term is zero only if no shift occurs in the long-run variance and the second only if a consistent in-sample estimate is obtained. However, the next four terms are zero on average, although the seventh possibly is not. This pattern then repeats, since the next block of four terms again is zero on average, with the penultimate term possibly non-zero, and the last zero on average. As with the earlier forecast error taxonomy, shifts in the mean seem pernicious, whereas those in the other parameters are much less serious contributors to forecast failure in variances. Indeed, even assuming a correct in-sample specification, so terms [2], [5], [7], [10], [12] all vanish, the main error components remain.

In practice, $\widehat{\varphi}_1 + \widehat{\varphi}_2$ is often close to unity, and $\widehat{\varphi}_0$ is small. This makes the behaviour of (53) also rather like a unit root in an AR(1) arising from unmodelled location shifts, even though the former remains non-integrated for constant parameters when the latter does not. In any case, models like (53) will miss jumps in volatility, but capture phases of quiescence and high volatility. Thus, consider forecasting using the variance equivalent of $\Delta^2 \widehat{x}_{T+1|T} = 0$, namely:

$$\widetilde{\sigma}_{T+1|T}^2 = \widehat{\sigma}_T^2. \tag{55}$$

Then (55) extrapolates the latest volatility estimate, and so will track the main changes in volatility, as well as constant variance periods, albeit noisely. All the earlier 'tricks' discussed above seem to apply again when the main focus is on variance forecasting (e.g., smoothed estimates of $\widehat{\sigma}_T^2$ etc.), as against interval forecasts, although related issues arise.

7 Conclusions

The properties of unpredictability of a random vector generated by a non-stationary process entail many of the difficulties that confront forecasting. Since econometric systems incorporate inter-temporal causal information representing inertial dynamics in the economy, they should have smaller prediction errors

than purely extrapolative devices—but in practice often do not. Rather, there are 10 basic difficulties to be circumvented to exploit any potential predictability, namely:
the composition of the DGP information set $\mathcal{I}_{\sqcup-\infty}$;
how $\mathcal{I}_{\sqcup-\infty}$ enters the DGP $\mathsf{D}_{\mathbf{y_t}}\left(\mathbf{y_t}|\mathcal{I}_{\sqcup-\infty}\right)$ (or for point forecasts, the form of the conditional expectation $\mathbf{f_t}\left(\mathcal{I}_{\sqcup-\infty}\right)$);
how $\mathsf{D}_{\mathbf{y_t}}\left(\mathbf{y_t}|\mathcal{I}_{\sqcup-\infty}\right)$ (or $\mathbf{f_t}\left(\mathcal{I}_{\sqcup-\infty}\right)$) changes over time;
the use of a limited information set $\mathcal{J}_{\sqcup-\infty} \subset \mathcal{I}_{\sqcup-\infty}$;
the mapping $\mathsf{D}_{\mathbf{y_t}}\left(\mathbf{y_t}|\mathcal{I}_{\sqcup-\infty}\right)$ into $\mathsf{D}_{\mathbf{y_t}}\left(\mathbf{y_t}|\mathcal{J}_{\sqcup-\infty}\right)$ inducing $\mathbf{g_t}\left(\mathcal{J}_{\sqcup-\infty}\right) = \mathsf{E_t}\left[\mathbf{f_t}\left(\mathcal{I}_{\sqcup-\infty}\right)|\mathcal{J}_{\sqcup-\infty}\right]$;
how $\mathcal{J}_{\mathcal{T}}$ will enter $\mathsf{D}_{\mathbf{y_{T+h}}}\left(\cdot|\mathcal{J}_{\mathcal{T}}\right)$ (or $\mathbf{g_{T+h}}\left(\mathcal{J}_{\mathcal{T}}\right)$) for a forecast origin at T;
approximating $\mathbf{g_t}\left(\mathcal{J}_{\sqcup-\infty}\right)$ by a function $\boldsymbol{\psi}\left(\mathcal{J}_{\sqcup-\infty}, \boldsymbol{\theta}\right)$ for some specification of the basic parameters $\boldsymbol{\theta}$;
measurement errors in $\widehat{\mathcal{J}}_{t-1}$ for $\mathcal{J}_{\sqcup-\infty}$;
the estimation of $\boldsymbol{\theta}$ from in-sample data $t = 1, \ldots, T$;
and the multistep nature of most economic forecasting.
The first six are aspects of predictability in the DGP; the second four of the formulation of forecasting models which seek to capture any predictability.

Two types of shift in $\mathbf{f_t}\left(\mathcal{I}_{\sqcup-\infty}\right)$ were distinguished, corresponding to mean-zero and location shifts respectively. The fundamental problem does not seem to be incomplete information *per se*: by construction, $\mathbf{g_t}\left(\mathcal{J}_{\sqcup-\infty}\right) - \mathbf{f_t}\left(\mathcal{I}_{\sqcup-\infty}\right)$ has a zero mean, even for processes with breaks. However, not knowing $\mathbf{g_t}\left(\mathcal{J}_{\sqcup-\infty}\right)$ is problematic for the specification of $\boldsymbol{\psi}\left(\mathcal{J}_{\sqcup-\infty}, \boldsymbol{\theta}\right)$ $\forall t$; the use of in-sample estimates when the process changes then compounds the difficulty.

Consequently, using a cointegrated linear dynamic system with breaks over the forecast horizon as the illustrative DGP, three adaptations were considered. The first was differencing the in-sample estimated DGP; the second was rapid updating of the estimated location in a growth representation; and the third was forecast-error correction mechanisms (FErCMs) loosely based on EWMAs. All three use representations that are knowingly mis-specified in-sample, and two use highly restricted choices of $\mathcal{J}_{\sqcup-\infty}$: nevertheless, they all help avoid systematic forecast failure. The analysis also highlighted the distinctly different role of the FErCM in EWMA (namely, to offset previous measurement errors) and in ICs (to offset breaks), which required the opposite sign. A synthesis in which the former role is combined with a different mechanism for adapting to location shifts has much to recommend it, and one univariate approach was noted.

The empirical example of the behaviour of M1 in the UK following the Banking Act of 1984 illustrated the three adaptations in action, with the last approximated by intercept corrections. All behaved as anticipated from the theory, and demonstrated the difficulty of out-performing 'naive extrapolative devices' when these are adative to location shifts that are inherently inimical to econometric systems. Overall, the outcomes suggest that, to retain causal information when the forecast-horizon 'goodness' of the model in use is unknown, model transformations may be the most reliable route of the three.

References

Allen, P. G., and Fildes, R. A. (2001). Econometric forecasting strategies and techniques. In Armstrong, J. S. (ed.), *Principles of Forecasting*, pp. 303–362. Boston: Kluwer Academic Publishers.

Bates, J. M., and Granger, C. W. J. (1969). The combination of forecasts. *Operations Research Quarterly*, **20**, 451–468. Reprinted in T.C. Mills (ed.), *Economic Forecasting*. Edward Elgar, 1999.

Bewley, R. A. (2000). Real-time forecasting with vector autoregressions: Spurious drift, structural

change and intercept-correction. Working paper, Economics Department, University of New South Wales, Sydney, Australia.

Bhansali, R. J. (1996). Asymptotically efficient autoregressive model selection for multistep prediction. *Annals of Institute of Statistical Mathematics*, **48**, 94–134.

Bhansali, R. J. (1997). Direct autoregressive predictors for multistep prediction: order selection and performance relative to the plug-in predictors. *Statistica Sinica*, **7**, 425–449.

Bhansali, R. J. (1999). Parameter estimation and model selection for multistep prediction of time series: a review. In Gosh, S. (ed.), *Asymptotics, Nonparametrics and Time Series*, pp. 201–225. New York, NY: Marcel Dekker.

Bollerslev, T. (1986). Generalised autoregressive conditional heteroskedasticity. *Journal of Econometrics*, **51**, 307–327.

Bontemps, C., and Mizon, G. E. (2003). Congruence and encompassing. In Stigum, B. P. (ed.), *Econometrics and the Philosophy of Economics*, pp. 354–378. Princeton: Princeton University Press.

Boswijk, H. P. (1992). *Cointegration, Identification and Exogeneity*, Vol. 37 of *Tinbergen Institute Research Series*. Amsterdam: Thesis Publishers.

Cairncross, A. (1969). Economic forecasting. *Economic Journal*, **79**, 797–812. Reprinted in T.C. Mills (ed.), *Economic Forecasting*. Edward Elgar, 1999.

Chevillon, G., and Hendry, D. F. (2002). Non-parametric direct multi-step estimation for forecasting economic processes. *International Journal of Forecasting*, **21**, 201–218.

Clemen, R. T. (1989). Combining forecasts: A review and annotated bibliography. *International Journal of Forecasting*, **5**, 559–583. Reprinted in T.C. Mills (ed.), *Economic Forecasting*. Edward Elgar, 1999.

Clements, M. P., and Hendry, D. F. (1993). On the limitations of comparing mean squared forecast errors (with discussion). *Journal of Forecasting*, **12**, 617–637. Reprinted in T.C. Mills (ed.), *Economic Forecasting*. Edward Elgar, 1999.

Clements, M. P., and Hendry, D. F. (1996). Forecasting in macro-economics. in Cox *et al.* (1996), pp. 101–141.

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

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

Clements, M. P., and Hendry, D. F. (2001a). Explaining the results of the M3 forecasting competition. *International Journal of Forecasting*, **17**, 550–554.

Clements, M. P., and Hendry, D. F. (2001b). An historical perspective on forecast errors. *National Institute Economic Review*, **177**, 100–112.

Clements, M. P., and Hendry, D. F. (2002). Modelling methodology and forecast failure. *The Econometrics Journal*, **5**, 319–344.

Clements, M. P., and Hendry, D. F. (1996). Multi-step estimation for forecasting. *Oxford Bulletin of Economics and Statistics*, **58**, 657–683.

Cox, D. R., Hinkley, D. V., and Barndorff-Nielsen, O. E. (eds.)(1996). *Time Series Models: In econometrics, finance and other fields*. London: Chapman and Hall.

Diebold, F. X., and Lopez, J. A. (1996). Forecast evaluation and combination. In Maddala, G. S., and Rao, C. R. (eds.), *Handbook of Statistics*, Vol. 14, pp. 241–268: Amsterdam: North–Holland.

Diebold, F. X., and Pauly, R. (1987). Structural change and the combination of forecasts. *Journal of Forecasting*, **6**, 21–40.

Engle, R. F. (1982). Autoregressive conditional heteroscedasticity, with estimates of the variance of United Kingdom inflation. *Econometrica*, **50**, 987–1007.

Fildes, R., and Ord, K. (2002). Forecasting competitions – their role in improving forecasting practice and research. In Clements, M. P., and Hendry, D. F. (eds.), *A Companion to Economic Forecasting*, pp. 322–253. Oxford: Blackwells.

Forni, M., Hallin, M., Lippi, M., and Reichlin, L. (2000). The generalized factor model: Identification and estimation. *Review of Economics and Statistics*, **82**, 540–554.

Granger, C. W. J. (1989). Combining forecasts - Twenty years later. *Journal of Forecasting*, **8**, 167–173.

Granger, C. W. J., and Pesaran, M. H. (2000a). A decision-theoretic approach to forecast evaluation. In Chon, W. S., Li, W. K., and Tong, H. (eds.), *Statistics and Finance: An Interface*, pp. 261–278. London: Imperial College Press.

Granger, C. W. J., and Pesaran, M. H. (2000b). Economic and statistical measures of forecasting accuracy. *Journal of Forecasting*, **19**, 537–560.

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

Harvey, A. C., and Shephard, N. G. (1992). Structural time series models. In Maddala, G. S., Rao, C. R., and Vinod, H. D. (eds.), *Handbook of Statistics*, Vol. 11. Amsterdam: North-Holland.

Hendry, D. F. (1979). Predictive failure and econometric modelling in macro-economics: The transactions demand for money. In Ormerod, P. (ed.), *Economic Modelling*, pp. 217–242. London: Heinemann. Reprinted in Hendry, D. F., *Econometrics: Alchemy or Science?* Oxford: Blackwell Publishers, 1993, and Oxford University Press, 2000; and in J. Campos, N.R. Ericsson and D.F. Hendry (eds.), *General to Specific Modelling*. Edward Elgar, 2005.

Hendry, D. F. (1995). *Dynamic Econometrics*. Oxford: Oxford University Press.

Hendry, D. F. (1997). The econometrics of macroeconomic forecasting. *Economic Journal*, **107**, 1330–1357. Reprinted in T.C. Mills (ed.), *Economic Forecasting*. Edward Elgar, 1999.

Hendry, D. F. (2000). On detectable and non-detectable structural change. *Structural Change and Economic Dynamics*, **11**, 45–65. Reprinted in *The Economics of Structural Change*, Hagemann, H. Landesman, M. and Scazzieri, R. (eds.), Edward Elgar, Cheltenham, 2002.

Hendry, D. F., and Clements, M. P. (2003). Economic forecasting: Some lessons from recent research. *Economic Modelling*, **20**, 301–329. European Central Bank, Working Paper 82.

Hendry, D. F., and Clements, M. P. (2004). Pooling of forecasts. *Econometrics Journal*, **7**, 1–31.

Hendry, D. F., and Doornik, J. A. (1994). Modelling linear dynamic econometric systems. *Scottish Journal of Political Economy*, **41**, 1–33.

Hendry, D. F., and Ericsson, N. R. (1991). Modeling the demand for narrow money in the United Kingdom and the United States. *European Economic Review*, **35**, 833–886.

Hendry, D. F., and Mizon, G. E. (1993). Evaluating dynamic econometric models by encompassing the VAR. In Phillips, P. C. B. (ed.), *Models, Methods and Applications of Econometrics*, pp. 272–300. Oxford: Basil Blackwell. Reprinted in J. Campos, N.R. Ericsson and D.F. Hendry (eds.), *General to Specific Modelling*. Edward Elgar, 2005.

Hendry, D. F., and Mizon, G. E. (2000). On selecting policy analysis models by forecast accuracy. In Atkinson, A. B., Glennerster, H., and Stern, N. (eds.), *Putting Economics to Work: Volume in Honour of Michio Morishima*, pp. 71–113. London School of Economics: STICERD.

Hendry, D. F., and Mizon, G. E. (2003). Forecasting in the presence of structural breaks and policy regime shifts. Forthcoming, D.W. Andrews, J.L. Powell, P.A. Ruud and J. Stock (eds.), *Identification and Inference in Econometric Models: Festschrift in Honour of T.J. Rothenberg*, Cambridge University Press, Cambridge.

Hendry, D. F., and von Ungern-Sternberg, T. (1981). Liquidity and inflation effects on consumers' expenditure. In Deaton, A. S. (ed.), *Essays in the Theory and Measurement of Consumers' Behaviour*, pp. 237–261. Cambridge: Cambridge University Press. Reprinted in Hendry, D. F., *Econometrics: Alchemy or Science?* Oxford: Blackwell Publishers, 1993, and Oxford University Press, 2000.

Johansen, S. (1992). A representation of vector autoregressive processes integrated of order 2. *Econometric Theory*, **8**, 188–202.

Kim, S., Shephard, N., and Chib, S. (1998). Stochastic volatility: Likelihood inference and comparison with ARCH models. *Review of Economic Studies*, **65**, 361–393.

Koopmans, T. C. (1937). *Linear Regression Analysis of Economic Time Series*. Haarlem: Netherlands Economic Institute.

Leitch, G., and Tanner, J. E. (1991). Economic forecast evaluation: Profits versus the conventional error measures. *American Economic Review*, **81**, 580–590. Reprinted in T.C. Mills (ed.), *Economic Forecasting*. Edward Elgar, 1999.

Makridakis, S., and Hibon, M. (2000). The M3-competition: Results, conclusions and implications. *International Journal of Forecasting*, **16**, 451–476.

Marget, A. W. (1929). Morgenstern on the methodology of economic forecasting. *Journal of Political Economy*, **37**, 312–339. Reprinted in Hendry, D. F. and Morgan, M. S. (1995), *The Foundations of Econometric Analysis*. Cambridge: Cambridge University Press; and in T.C. Mills (ed.), *Economic Forecasting*. Edward Elgar, 1999.

Melino, A., and Turnbull, S. M. (1990). Pricing foreign currency options with stochastic volatility. *Journal of Econometrics*, **45**, 239–265.

Mills, T. C. (ed.)(1999). *Economic Forecasting*. Cheltenham, UK: Edward Elgar. 2 vols.

Morgenstern, O. (1928). *Wirtschaftsprognose: eine Untersuchung ihrer Voraussetzungen und Möglichkeiten*. Vienna: Julius Springer.

Newbold, P., and Harvey, D. I. (2002). Forecasting combination and encompassing. In Clements, M. P., and Hendry, D. F. (eds.), *A Companion to Economic Forecasting*, pp. 268–283: Oxford: Blackwells.

Paruolo, P. (1996). On the determination of integration indices in I(2) systems. *Journal of Econometrics*, **72**, 313–356.

Persons, W. M. (1924). *The Problem of Business Forecasting*. No. 6 in Pollak Foundation for Economic Research Publications. London: Pitman.

Rahbek, A., Kongsted, H. C., and Jørgensen, C. (1999). Trend-stationarity in the I(2) cointegration model. *Journal of Econometrics*, **90**, 265–289.

Schweppe, F. (1965). Evaluation of likelihood functions for Gaussian signals. *IEEE Transactions on Information Theory*, **11**, 61–70.

Shephard, N. (1996). Statistical aspects of ARCH and stochastic volatility. in Cox *et al.* (1996), pp. 1–67.

Stock, J. H., and Watson, M. W. (1999). A comparison of linear and nonlinear models for forecasting

macroeconomic time series. In Engle, R. F., and White, H. (eds.), *Cointegration, Causality and Forecasting*, pp. 1–44. Oxford: Oxford University Press.

Tay, A. S., and Wallis, K. F. (2000). Density forecasting: A survey. *Journal of Forecasting*, **19**, 235–254. Reprinted in Clements, M. P. and Hendry, D. F. (eds.) *A Companion to Economic Forecasting*, pp.45 – 68, Oxford: Blackwells, 2002.

Taylor, S. J. (1986). *Modelling Financial Time Series*. Chichester: John Wiley.