Forecasting economic and financial time-series with non-linear models

Michael P. Clements\textsuperscript{a}, Philip Hans Franses\textsuperscript{b}, Norman R. Swanson\textsuperscript{c,*}

\textsuperscript{a}Department of Economics, University of Warwick, Warwick, RI, USA
\textsuperscript{b}Econometric Institute, Erasmus University Rotterdam, Rotterdam, NY, USA
\textsuperscript{c}Department of Economics, Rutgers University, 75 Hamilton Street, New Brunswick, NJ 08901-1248, USA

Abstract

In this paper we discuss the current state-of-the-art in estimating, evaluating, and selecting among non-linear forecasting models for economic and financial time series. We review theoretical and empirical issues, including predictive density, interval and point evaluation and model selection, loss functions, data-mining, and aggregation. In addition, we argue that although the evidence in favor of constructing forecasts using non-linear models is rather sparse, there is reason to be optimistic. However, much remains to be done. Finally, we outline a variety of topics for future research, and discuss a number of areas which have received considerable attention in the recent literature, but where many questions remain.

© 2004 International Institute of Forecasters. Published by Elsevier B.V. All rights reserved.

Keywords: Economic; Financial; Non-linear models

1. Introduction

Whilst non-linear models are often used for a variety of purposes, one of their prime uses is for forecasting, and it is in terms of their forecasting performance that they are most often judged. However, a casual review of the literature suggests that often the forecasting performance of such models is not particularly good. Some studies find in favor, but equally there are studies in which their added complexity relative to rival linear models does not result in the expected gains in terms of forecast accuracy. Just over a decade ago, in their review of non-linear time series models, De Gooijer and Kumar (1992) concluded that there was no clear evidence in favor of non-linear over linear models in terms of forecast performance, and we suspect that the situation has not changed very much since then. It seems that we have not come very far in the area of non-linear forecast model construction.

We argue that the relatively poor forecasting performance of non-linear models calls for substantive further research in this area, given that one might feel uncomfortable asserting that non-linearities are unimportant in describing economic and financial phenomena. The problem may simply be that our non-linear models are not mimicking reality any better than simpler linear approximations, and in the next section we discuss this and related reasons why a good forecast performance 'across the board' may constitute something of a 'holy grail' for non-linear models.

We discuss the current state-of-the-art in non-linear modeling and forecasting, with particular emphasis...
placed on outlining a number of open issues. The topics we focus on include joint and conditional predictive density evaluation, loss functions, estimation and specification, and data-mining, amongst others. As such, this paper complements the rest of the papers in this special issue of the *International Journal of Forecasting*.

The rest of the paper is organized as follows. In Section 2 we discuss why one might want to consider non-linear models, and a number of reasons why their forecasting ability relative to linear models may not be as good as expected. In Section 3 we discuss recent theoretical and methodological issues to do with forecasting with non-linear models, many of which go beyond the traditional preoccupation with point forecasts to consider the whole predictive density. Section 4 highlights a number of empirical issues, and how these are dealt with in the papers collected in this issue. Concluding remarks are gathered in Section 5.

### 2. Why consider non-linear models?

Many of us believe that linear models ought to be a relatively poor way of capturing certain types of economic behavior, or economic performance, at certain times. The obvious example would be a linear (e.g. Box–Jenkins ARMA) model of output growth in a Western economy subject to the business cycle, where the properties of output growth in recessions are in some ways quite different from expansions (e.g. Hamilton, 1989; Sichel, 1994, but references are innumerable). Output growth non-linearities can be characterized by the presence of two or more regimes (e.g. recessions and expansions), as can financial variables (periods of high and low volatility). Other types of non-linearity might include the possibility that the effects of shocks accumulate until a process “explodes” (self-exciting or catastrophic behavior), as well as the notion that some variables are relevant for forecasting only once in a while (e.g. only when oil prices increase by a large amount do they have a significant effect on output growth, and therefore, become useful for forecasting output growth).

In macroeconomics and finance theory a host of non-linear models are already in vogue. For example, almost all-real business cycle (RBC) models are highly non-linear. As well as which, bond pricing models, diffusion processes describing yield curves, and almost all other continuous time finance models are non-linear. The predominance of non-linear models in economics and finance is not inconsistent with the use of linear models by the applied practitioner, as such models can be viewed as reasonable approximations to the non-linear phenomenon of interest. Thus, from the perspective of forecasting, there is ample reason to continue to look at non-linear models. As our non-linear model estimation, selection and testing approaches become more sophisticated, one might expect to see their forecast performance improve commensurately. It is thus not surprising that non-linear models, ranging from regime-switching models, to neural networks and genetic algorithms, are receiving a great deal of attention in the literature.

On a more cautionary note, we review a number of factors, which might count against the aforementioned improvement in the relative performance of non-linear models. Shortly after De Gooijer and Kumar, Granger and Teräsvirta (1993a); Granger and Teräsvirta (1993b), ch. 9 (see also Teräsvirta & Anderson, 1992) in their review of smooth transition autoregressive (STAR) models of US industrial production, argue that the superior in-sample performance of such models will only be matched out-of-sample if that period contains ‘non-linear features’. Similarly, Tong (1995), pp. 409–410 argues strongly that for non-linear models ‘how well we can predict depends on where we are’ and that there are ‘windows of opportunity for substantial reduction in prediction errors’. This suggests that an important aspect of an evaluation of the forecasts from non-linear models relative to the linear AR models is to make the comparison in a way which highlights the favorable performance of the former for certain states, especially if it is forecasts in those states which are most valuable to the user of the forecasts. Clements and Smith (1999) compare the forecasting performance of empirical self-exciting threshold autoregressive (SETAR) models and AR models using simulation techniques which ensure that past non-linearities are present in the forecast period. See Tiao and Tsay (1994) for a four-regime TAR model applied to US GNP, and Boero and Marrocu (2004) for an application of regime-specific evaluation to exchange rate forecasts.
In addition, in the context of exchange rate prediction, Diebold and Nason (1990) give a number of reasons why non-linear models may fail to outperform linear models. One is that apparent non-linearities detected by tests for linearity are due to outliers or structural breaks, which cannot be readily exploited to improve out-of-sample performance, and may only be detected by careful analysis along the lines of Koop and Potter (2000), for example. They also suggest that conditional—mean non-linearities may be a feature of the data generating process (DGP), but may not be large enough to yield much of an improvement to forecasting, as well as the explanation that they are present and important, but that the wrong types of non-linear models have been used to try and capture them.

There is a view that, because some aspects of the economy or financial markets do indeed display non-linear behavior, then neglecting these features in constructing forecasts would leave the end-user uneasy, feeling that the forecasts are in some sense “second”. This follows from the belief that a good model for the in-sample data should also be a good out-of-sample forecasting model. As an example, an AR(2) model for US GNP growth may yield lower average squared errors than an artificial neural network with one hidden unit, but, knowing that the AR(2) model is incapable of capturing the distinct dynamics of the business cycle phases, does not necessarily translate in to a good out-of-sample performance relative to a model such as an AR. The sub-section below illustrates, and see also Clements and Hendry (1999) for a more general discussion.

We take the view that, if one believes the underlying phenomenon is non-linear, it is worth considering a non-linear model, but warn against the expectation that such models will always do well—there are too many unknowns and the economic system is too complex to support the belief that simply generalizing a linear model in one (simple!) direction, such as adding another regime, will necessarily improve matters. That said, it is natural to be unhappy with models, which are obviously deficient in some respect, and to seek alternatives. The subsequent sections of this paper review some of the recent developments in model selection and empirical strategies, as well as the remaining papers in this issue, which take up the challenge of forecasting with non-linear models.

We end this section with two short illustrations of some of the difficulties that can arise. In the first, there is distinct regime-switching behavior, but this does not contribute to an improved forecast performance. In the second, we discuss the relationship between output growth and the oil price.

2.1. Markov-switching models of US output growth

Clements and Krolzig (1998) present some theoretical explanations for why Markov-switching (MS) models may not forecast much better than AR models, and apply their analysis to post war US output growth. The focus is on a two-regime MS model:

$$\Delta y_t - (s_t) = \alpha (\Delta y_{t-1} - (s_{t-1})) + u_t,$$

where $u_t \sim IN[0, \sigma_u^2]$. The conditional mean $\mu(s_t)$ switches between two states:

$$\mu(s_t) = \begin{cases} 
\mu_1 > 0 & \text{if } s_t = 1 (\text{"expansion" or "boom"}), \\
\mu_2 < 0 & \text{if } s_t = 2 (\text{"contraction" or "recession"}),
\end{cases}$$

(2)

The description of a MS–AR model is completed by the specification of a model for the stochastic and unobservable regimes on which the parameters of the conditional process depend. Once a law has been specified for the states $s_t$, the evolution of regimes can be inferred from the data. The regime-generating process is assumed to be an ergodic Markov chain with a finite number of states $s_t = 1, 2$, (for a two-regime model), defined by the transition probabilities:

$$p_{ij} = \Pr(s_{t+1} = j \mid s_t = i),$$

$$\sum_{j=1}^{2} p_{ij} = 1 \quad \forall i, j \in \{1, 2\}. \quad (3)$$

The model can be rewritten as the sum of two independent processes:

$$\Delta y_t - \mu_1 = \mu_t + z_t,$$
where \( \mu_t \) is the unconditional mean of \( \Delta y_t \), such that \( E[\mu_t] = E[z_t] = 0 \). While the process \( z_t \) is Gaussian:

\[
z_t = \alpha z_{t-1} + \varepsilon_t, \quad \varepsilon_t \sim \mathcal{N}(0, \sigma^2_\varepsilon),
\]

the other component, \( \mu_t \), represents the contribution of the Markov chain:

\[
\mu_t = (\mu_2 - \mu_1) \xi_t,
\]

where \( \xi_t = Pr(s_t = 2) \) if \( s_t = 2 \) and \( -Pr(s_t = 2) \) otherwise. \( Pr(s_t = 2) = p_{12}/(p_{11} + p_{21}) \) is the unconditional probability of regime 2. Invoking the unrestricted VAR(1) representation of a Markov chain:

\[
\xi_t = (p_{11} + p_{22} - 1) \xi_{t-1} + \eta_t,
\]

then predictions of the hidden Markov chain are given by:

\[
\hat{\xi}_{T+h | T} = (p_{11} + p_{22} - 1)^h \hat{\xi}_{T | T}
\]

where \( \hat{\xi}_{T | T} = E[\xi_T \mid Y_T] = Pr(s_T = 2 \mid Y_T) - Pr(s_T = 2) \) is the filtered probability of being in regime 2 corrected for the unconditional probability. Thus, the conditional mean of \( \Delta y_{T+h} \) is given by \( \hat{\Delta} y_{T+h | T} = \mu \), which equals:

\[
\hat{\xi}_{T+h | T} = \mu \]

The first term in (4) is the optimal prediction rule for a linear model, and the contribution of the Markov regime-switching structure is given by the term multiplied by \( \hat{\xi}_{TAT} \), where \( \hat{\xi}_{TAT} \) contains the information about the most recent regime at the time the forecast is made. Thus, the contribution of the non-linear part of (4) to the overall forecast depends on both the magnitude of the regime shifts, \( |\mu_2 - \mu_1| \), and on the persistence of regime shifts \( p_{11} + p_{22} - 1 \) relative to the persistence of the Gaussian process, given by \( \alpha \).

Clements and Krolzig estimate \( p_{11} + p_{22} - 1 = 0.65 \), and the largest root of the AR polynomial to be 0.64, so that the second reason explains the success of the linear AR model in forecasting. Since the predictive power of detected regime shifts is extremely small, \( p_{11} + p_{22} - 1 = \alpha \) in (4), the conditional expectation collapses to a linear prediction rule. Heuristically, the relative performance of non-linear regime-switching models would be expected to be better the more persistent the regimes. When the regimes are unpredictable, we can do no better than employing a simple linear model. See also Krolzig (2003); Dacco and Satchell (1999) for an analysis of the effects of wrongly classifying the regime the process will be in.

A number of authors, such as Sensier, Artis, Osborn and Birchenhall (2004), attempt to predict business cycle regimes, and the transition probabilities in (4) can be made to depend on leading indicator variables, as a way of sharpening the forecasting ability of these models. Franses, Paap and Vroomen (2004) utilize information from extraneous variables to determine the regime in a novel approach to predicting the US unemployment rate.

### 2.2. Output growth and the oil price

Of obvious interest in the literature is the relationship between oil prices and the macroeconomy. To what extent did the OPEC oil price rises contribute to the recessions in the 1970s, and might one expect reductions in prices to stimulate growth, perhaps to a lesser extent? Hamilton (1983) originally proposed a linear relationship between oil prices and output growth for the US. This was challenged by Mork (1989), who suggested that the relationship was asymmetric, in that output growth responds negatively to oil price increases, but is unaffected by oil price declines. With the advantage of several more years of data, Hooker (1996) showed that the linear relationship proposed by Hamilton (1983) appears not to hold from 1973 onwards (the date of the first oil price hike!). However, he also cast doubt on the simple asymmetry hypothesis suggested by Mork (1989).

More recently, Hamilton (1996) proposes relating output growth to the net increase in oil prices over the previous year, and constructs a variable that is the percentage change in the oil price in the current quarter over the previous year’s high, when this is positive, and otherwise takes on the value zero. Thus, increases in the price of oil, which simply reverse previous (within the preceding year) declines do not depress output growth. Recently, Hamilton (2000) has used a new flexible non-linear approach (Dahl &
Raymond and Rich (1997) have investigated the relationship between oil prices and the macroeconomy by including the net increase in the oil price in an MS model of US output for the period 1951–1995. They are interested in whether the recurrent shifts between expansion and contraction identified by the MS model remain when oil prices have been included as an explanatory variable for the mean of output. Raymond and Rich (1997) conclude that ‘while the behavior of oil prices has been a contributing factor to the mean of low growth phases of output, movements in oil prices generally have not been a principal determinant in the historical incidence of these phases . . .’ (p. 196). Further, Clements and Krolzig (2002) investigate whether oil prices can account for the asymmetry in the business cycle using the tests proposed in Clements and Krolzig (2003).

Clearly, the process of discovery of (an approximation to) the form of the non-linearity in the relationship between output growth and oil price changes has taken place over two decades, is far from simple, and has involved the application of state of the art econometric techniques. Perhaps this warns against expecting too much in the near future using ‘canned’ routines and models.

3. Theoretical and methodological issues

There are various theoretical issues involved in constructing non-linear models for forecasting. In this section we outline a number of these. An obvious starting point is which non-linear model to use, given the many possibilities that are available, even once we have determined the purpose to which it is to be put (here, forecasting). The different types of models often require different theoretical and empirical tools (see e.g. the recent surveys by Franses & van Dijk, 2000; van Dijk, Franses & Teräsvirta, 2002). For example, closed form solutions exist for the conditional mean forecast for an MS process, but not for a threshold autoregressive process, requiring simulation or numerical methods in the latter case. Certain theoretical properties, such as stability and stationarity, and the persistence of shocks, are not always immediately evident.

The choice of model might be suggested by economic theory, and often by the requirement that the model is capable of generating the key characteristics of the data at hand. Of course, the issue of which characteristics often arises. For example, Pagan (1997a); Pagan (1997b); Harding and Pagan (2001) argue that non-linear models should be evaluated in terms of their ability to reproduce certain features of the classical cycle, rather than their ability to match the stationary moments of the detrended growth cycle.1

An approach to tackling these issues is to use predictive density or distributional testing, as a means of establishing which of a number of candidate forecasting models has distributional features that most closely match the historical record. This could include, for example, finding out which of the models yields the best distributional or interval predictions. For example, in financial risk management interest often focuses on predicting a particular quantile (as the Value-at-Risk, VaR) but alternatively the entire conditional distribution of a variable may be of interest. Over the last few years, a new strand of literature addressing the issue of predictive density evaluation has arisen (see e.g. Bai, 2001; Christoffersen, 1998; Christoffersen, Hahn & Inoue, 2001; Clements & Smith, 2000, 2002; Diebold, Gunther & Tay, 1998 (henceforth DGT), Diebold, Hahn & Tay, 1999; Giacomini & White, 2003; Hong, 2001). The literature on the evaluation of predictive densities is largely concerned with testing the null of correct dynamic specification of an individual conditional distribution model. At the same time, the point forecast evaluation literature explicitly recognizes that all the candidate models may be misspecified (see e.g. Corradi & Swanson, 2002; White, 2000). Corradi and Swanson (2003a) draw on elements from both types of papers in order to provide a test for choosing among competing predictive density models which may be misspecified. Giacomini and White (2003)

---

1 As an aside, these papers have shown that the durations and amplitudes of expansions and contractions of the classical cycle can be reasonably well reproduced by simple random walk with drift models, where the ratio of the drift to the variance of the disturbance term is the crucial quantity. Non-linear models appear to add little over and above that which can be explained by the random walk with drift.
tackle a similar problem, developing a framework that allows for the evaluation of both nested and nonnested models.

Many of the above papers seek to evaluate predictive densities by testing whether they have the property of correct (dynamic) specification. By making use of the probability integral transform, DGT suggest a simple and effective means by which predictive densities can be evaluated. Using the DGT terminology, if \( p_t(y_t | \Omega_{t-1}) \) is the “true” conditional distribution of \( y_t \), then the probability of observing a value of \( y_t \) no larger than that actually observed is a uniform random variable on [0, 1]. Moreover, if we have a sequence of predictive densities, then the resulting sequence of probabilities should be identically independently distributed. Goodness-of-fit statistics can then be constructed that compare the empirical distribution function of the probabilities to the 45° degree line, possibly taking into account that

\[
\Pr(y_{t+1} \leq u | Z_t) = Pr(q_i(y_t, Z_t^{-1}, \theta_i)) = \frac{1}{C_0} \int_U E((F_i(u | Z_t', \theta_i^1) - F_0(u | Z_t', \theta_0))^2) \phi(u) du \leq 0
\]

where \( q_i \) denotes the objective function for model \( i \). Following standard practice (such as in the real-time forecasting literature), this estimator is first computed using \( R \) observations, then \( R + 1 \) observations, then \( R + 2 \), and so on until the last estimator is constructed using \( T - 1 \) observations, resulting in a sequence of \( P = T - R \) estimators. In the current discussion, we focus on 1-step ahead prediction, so these estimators are then used to construct sequences of \( P \) 1-step ahead forecasts and associated forecast errors.

Now, define the group of conditional distribution models from which we want to make a selection as \( F_1(u | Z_t', \theta_1), \ldots, F_n(u | Z_t', \theta_n) \), and define the true conditional distribution as \( F_0(u | Z_t', \theta_0) = Pr(y_{t+1} \leq u | Z_t) \). Hereafter, assume that \( q_i(y_t, Z_t^{-1}, \theta_i) = -\ln f_i(y_t | Z_t^{-1}, \theta_i) \), where \( f_i(\cdot |, \theta) \) is the conditional density associated with \( F_i, i = 1, \ldots, n \), so that \( \theta_i \) is the probability limit of a quasi maximum likelihood estimator (QMLE). If model \( i \) is correctly specified, then \( \theta_i = \theta_0 \). Now, \( F_i(\cdot |, \theta_i) \) is taken as the benchmark model, and the objective is to test whether some competitor model can provide a more accurate approximation of \( F_0(\cdot |, \theta_0) \) than the benchmark. Assume that accuracy is measured using a distributional analog of mean square error. More precisely, the squared (approximation) error associated with model \( i, i = 1, \ldots, n \), is measured in terms of the average over \( U \) of \( E((F_i(u | Z_t', \theta_i^1) - F_0(u | Z_t', \theta_0))^2) \), where \( u \in U \), and \( U \) is a possibly unbounded set on the real line. The hypotheses of interest are:

\[
H_0 : \max_{k=2, \ldots, n} \int_U E((F_i(u | Z_t', \theta_i^1) - F_0(u | Z_t', \theta_0))^2 - (F_k(u | Z_t', \theta_k^1) - F_0(u | Z_t', \theta_0))^2) \phi(u) du \leq 0
\]

versus

\[
H_A : \max_{k=2, \ldots, n} \int_U E((F_i(u | Z_t', \theta_i^1) - F_0(u | Z_t', \theta_0))^2 - (F_k(u | Z_t', \theta_k^1) - F_0(u | Z_t', \theta_0))^2) \phi(u) du > 0,
\]

where \( \phi(u) \geq 0 \) and \( \int_U \phi(u) = 1 \), \( u \in U \subset \mathbb{R} \), \( U \) possibly unbounded. Note that for a given \( u \), we compare
conditional distributions in terms of their (mean square) distance from the true distribution. The statistic is:

\[ Z_P = \max_{k=2, \ldots, n} \int_U Z_{P,n}(1, k) \phi(u)du, \tag{9} \]

where

\[ Z_{P,n}(1, k) = \frac{1}{\sqrt{P}} \sum_{i=R}^{T-1} \left[ (1\{y_{i+1} \leq u\} - F_1(u | Z^t, \hat{\theta}_{1,t})^2 \right. \]

\[ \left. - (1\{y_{i+1} \leq u\} - F_k(u | Z^t, \hat{\theta}_{k,t})^2) \right]. \tag{10} \]

Here, each model is estimated via QMLE, so that in terms of the above notation, \( q_i = -\ln f_i \), where \( f_i \) is the conditional density associated with model \( i \), and \( \hat{\theta}_{i,t} \) is defined as \( \hat{\theta}_{i,t} = \arg \max_{\theta_i \in \Theta_{i,T}} \frac{1}{T} \sum_{j=1}^{T} \ln f_j(y_j, Z_t, \theta_i) \), \( R = t \leq T - 1 \), \( i = 1, \ldots, n \). For further details, please refer to Corradi and Swanson (2003a).

Clearly, the above approach can be used to evaluate multiple non-linear forecasting models. Now, assume that focus centers on evaluating the joint dynamics of multiple non-linear forecasting models. Now, assume that the variables of interest are output and lagged output. Now, set model 1 as the benchmark model. Let \( \Delta \log X_t, t = 1, \ldots, T \) denote actual historical output (growth rates), and let \( \Delta \log X_{j,n}, j = 1, \ldots, m \) and \( n = 1, \ldots, S \). denote the output series simulated under model \( j \), where \( S \) denotes the length of the simulated sample. Denote \( \Delta \log X_{j,n}(\hat{\theta}_{j,T}), n = 1, \ldots, S, j = 1, \ldots, m \) to be a sample of length \( S \) drawn (simulated) from model \( j \) and evaluated at the parameters estimated under model \( j \), where parameter estimation is done using the \( T \) available historical observations. Further, let \( Y_t = (\Delta \log X_t, \Delta \log X_{t-1}, \Delta \log X_{j,n}(\hat{\theta}_{j,T}), \Delta \log X_{j,n-1}(\hat{\theta}_{j,T})), \) and let \( F_0(u; \theta_0) \) denote the distribution of \( Y_t \) evaluated at \( u \) and \( F_j(u; \theta_j) \) denote the distribution of \( Y_t(u; \theta_j) \), where \( \theta_j \) is the probability limit of \( \hat{\theta}_{j,T} \), taken as \( T \rightarrow \infty \), and where \( u \in \mathcal{U} \subset \mathbb{R}^2 \), possibly unbounded. As above, accuracy is measured in terms of squared error. The squared (approximation) error associated with model \( j \), \( j = 1, \ldots, m \), is measured in terms of the (weighted) average over \( U \) of \( (F(u; \theta_j) - F_0(u; \theta_0))^2 \), where \( u \in \mathcal{U} \), and \( U \) is a possibly unbounded set on \( \mathbb{R}^2 \). Thus, the rule is to choose Model 1 over Model 2 if

\[ \int_U (F_1(u; \theta_1^\dagger) - F_0(u; \theta_0))^2 \phi(u)du < \int_U ((F_2(u; \theta_2^\dagger) - F_0(u; \theta_0))^2 \phi(u)du, \]

where \( \int_U \phi(u)du = 1 \) and \( \phi(u) \geq 0 \) for all \( u \in \mathcal{U} \subset \mathbb{R}^2 \). For any evaluation point, this measure defines a norm and is a typical goodness-of-fit measure. Note that within our context, the hypotheses of interest are:

\[ H_0 : \max_{j=2, \ldots, m} \int_U ((F_0(u; \theta_0) - F_1(u; \theta_1^\dagger))^2 \]

\[ - (F_0(u) - F_j(u; \theta_j^\dagger))^2 \phi(u)du \leq 0 \]
and

\[ H_A : \max_{j=2,\ldots,m} \int_U \left( (F_0(u) - F_1(u; \theta_1^j))^2 - (F_0(u) - F_j(u; \theta_1^j))^2 \right) \phi(u) du > 0. \]

Thus, under \( H_0 \), no model can provide a better approximation (in a squared error sense) to the distribution of \( Y \) than the approximation provided by model 1. If interest focuses on confidence intervals, so that the objective is to “approximate” \( \Pr(\hat{u} \leq Y \leq \bar{u}) \) then the null and alternative hypotheses can be stated as:

\[ H_0^0 : \max_{j=2,\ldots,m} \left( (F_1(\bar{u}; \theta_1^j) - F_j(\bar{u}; \theta_1^j)) - (F_0(\bar{u}; \theta_0) - F_0(\bar{u}; \theta_0)) \right)^2 \leq 0. \]

versus

\[ H_A^0 : \max_{j=2,\ldots,m} \left( (F_1(\bar{u}; \theta_1^j) - F_1(\bar{u}; \theta_1^j)) - (F_0(\bar{u}; \theta_0) - F_0(\bar{u}; \theta_0)) \right)^2 > 0. \]

If interest focuses on testing the null of equal accuracy of two distribution models (analogous to the pairwise conditional mean comparison setup of Diebold & Mariano, 1995), we can simply state the hypotheses as:

\[ H_0^1 : \int_U \left( (F_0(u; \theta_0) - F_1(u; \theta_1^j))^2 - (F_0(u; \theta_0) - F_0(u; \theta_0))^2 \right) \phi(u) du = 0 \]

versus

\[ H_A^1 : \int_U \left( (F_0(u; \theta_0) - F_1(u; \theta_1^j))^2 - (F_0(u; \theta_0) - F_0(u; \theta_0))^2 \right) \phi(u) du \neq 0, \]

In order to test \( H_0 \) versus \( H_A \), the relevant test statistic is \( \sqrt{T}Z_{TS} \) where:

\[ Z_{TS} = \max_{j=2,\ldots,m} \int_U Z_{j,T,S}(u) \phi(u) du, \quad (11) \]

with \( \hat{\theta}_{j,T} \) an estimator of \( \theta_j^j \) that satisfies Assumption 2 below. See Corradi and Swanson (2003c) for further details.

Another measure of distributional accuracy available in the literature (see e.g. White, 1982; Vuong, 1989), is the KLIC, according to which we should choose Model 1 over Model 2 if:

\[ E(\log f_1(Y; \theta_1^j) - \log f_2(Y; \theta_1^j)) > 0. \]

The KLIC is a sensible measure of accuracy, as it chooses the model, which on average gives higher probability to events, which have actually occurred. Also, it leads to simple Likelihood Ratio tests. Interestingly, Fernandez-Villaverde and Rubio-Ramirez (2001) have shown that the best model under the KLIC is also the model with the highest posterior probability. The above approach is an alternative to the KLIC that should be viewed as complementary in some cases, and preferred in others. For example, if we are interested in measuring accuracy over a specific region, or in measuring accuracy for a given confidence interval, this cannot be done in an obvious manner using the KLIC, while it can easily be done using our measure. As an illustration, assume that we wish to evaluate the accuracy of different models in approximating the probability that the rate of growth of output is say between 0.5% and 1.5%. We can do so quite easily using the squared error criterion, but not using the KLIC. Furthermore, we often do not have an explicit form for the density implied by the various models we are comparing. Of course, model comparison can be done using kernel density estima-

---

2 \( H_0^1 \) versus \( H_A^1 \) and \( H_0^0 \) versus \( H_A^0 \) can be tested in a similar manner.

3 Recently, Giacomini (2002) proposes an extension, which uses a weighted (over \( Y \)) version of the KLIC, and Kitamura (2002) suggests a generalization for choosing among models that satisfy some conditional moment restrictions.
tors, within the KLIC framework. However, this leads to tests with non-parametric rates (see e.g. Zheng, 2000). On the other hand, comparison via our squared error measure of accuracy is carried out using empirical distributions, so that resulting test statistics converge at parametric rates.

Point forecast production and evaluation continues to receive considerable attention, and can perhaps be viewed as a leading indicator for the predictive density literature. There are now a host of tests based on traditional squared error loss criteria, but in addition tests based on directional forecast accuracy and sign tests, tests of forecast encompassing, as well as measures and tests based on other loss functions. Swanson and White (1997a) survey a number of the important contributions, which include Chao, Corradi and Swanson (2001); Chatfield (1993); Clark and McCracken (2001); Clements and Hendry (1993); Diebold and Chen (1996); Diebold and Mariano (1995); Hansen (2001); Hansen, Heaton and Luttmert (1995); Hansen and Jeganathan (1997); Harvey, Leybourne and Newbold (1997); Harvey, Leybourne and Newbold (1998); Linton, Maasoumi and Whang (2003); McCracken (1999); Pesaran and Timmerman (1992); Pesaran and Timmerman (1994); Pesaran and Timmerman (2000); Stekler (1991); Stekler (1994); West (1996); West (2001); West and McCracken (2002), among others. A number of these papers emphasise the role of parameter estimation uncertainty in testing for equal forecast accuracy, or that one model forecast encompasses another, as well as the consequences of the models being nested. Nevertheless, much remains to be done with regard to making these tests applicable to non-linear models, and constructing tests using non-linear and/or non-differentiable loss functions, as well as allowing for parameter estimation error and misspecification when the comparisons involve non-linear models.

The above paragraph notes the role of the loss function in determining how accuracy is to be assessed. However, there is also the issue of whether the in-sample model estimation criterion and out-of-sample forecast accuracy criterion should be matched. For example, it is only recently that much attention has been given to the notion that the same loss function used in-sample for parameter estimation is often that which should be used out-of-sample for forecast evaluation. Much remains to be done in this area, although progress has been made, as discussed in Christoffersen and Diebold (1996); Christoffersen and Diebold (1997); Clements and Hendry (1996); Granger (1993); Granger (1999); Weiss (1996). That said, it is often difficult to come up with asymptotically valid inferential strategies using standard estimation procedures (that essentially minimize one-step ahead errors) for many varieties of non-linear models, from smooth transition models to projection pursuit and wavelet models. In some contexts it is even difficult to establish the consistency of some econometric parameter estimates, such as cointegrating vectors in certain non-linear cointegration models, and threshold parameters in some types of regime-switching models. In sum, estimation and in-sample inference of non-linear forecasting models remains a potentially difficult task, with much work remaining to be done. Nevertheless, there are many recent papers that propose novel approaches to estimation, such as the variety of new cross-validation related techniques in the area of neural nets.

A general problem with non-linear models is the ‘curse of dimensionality’ and the fact that such models tend to have a large number of parameters (at least relative to the available number of macro-economic data points)—how to keep the number of parameters at a tractable level? This sort of issue is relevant to the specification of many varieties of non-linear models, including smooth transition models (see e.g. Granger & Teräsvirta, 1993a,b) and neural network models (see e.g. Swanson & White, 1995, 1997a,b). A counter to the fear that such models may be overfitting in-sample—in the sense of picking up transient, accidental connections between variables—is of course to compare the models on out-of-sample performance, or on a ‘hold-out’ sample. The ‘data-snooping procedures’ developed by White (2000), and used in Sullivan, Timmerman and White (1999); Sullivan, Timmerman and White (2001); Sullivan, Timmerman and White (2003), can also be used, but this would appear to be an open area, with much room for advance. For example, the extension of the data-snooping methodology to multivariate models awaits attention, where the sheer magnitude of the problem can quickly grow out of hand.

Finally, although not germane to forecasting, some models have parameters that are readily interpretable, whilst others are less clear. The study by De Gooijer
and Vidiella-i-Anguera (2004) extends ‘non-linear cointegration’ to allow the equilibrium, cointegrating relationship to depend upon the regime. Difficult issues arise when cointegration is no longer ‘global’, and the theory-justification for the long-run relationship is less clear. Clements and Galvão (2004) review specification and estimation procedures in systems when cointegration is ‘global’, and evaluate the forecasting ability of non-linear systems in the context of interest rate prediction, building on earlier contributions by Anderson (1997); van Dijk and Franses (2000); Kunst (1992), inter alia. Using a variety of forecast evaluation methods, De Gooijer and Vidiella-i-Anguera establish the superiority of their model from a forecasting perspective.

Another new model is developed by Franses, Paap and Vroomen (2004), who propose a model in which a key autoregressive parameter depends on a leading indicator variable. This model captures the notion that some variables are relevant for forecasting only once in a while, and it mimics some of the ideas put forward in Franses and Paap (2002). The autoregressive parameter is constant unless a linear function of the leading indicator plus a disturbance term exceeds a certain threshold level. The authors discuss issues relating to estimation, inference and forecasting, and the relationship of their model to existing non-linear models. The model is applied to forecasting unemployment, and is shown to be capable of capturing the sharp increases in unemployment in recessions, and to provide competitive forecasts compared to alternative models.

A number of possible approaches are available to account for the possibility that the parameters in forecasting models are changing over time. Taylor (2004) uses adaptive exponential smoothing methods that allow smoothing parameters to change over time, in order to adapt to changes in the characteristics of the time series. More specifically, he presents a new adaptive method for predicting the volatility in financial returns, where the smoothing parameter varies as a logistic function of user-specified variables. The approach is analogous to that used to model time-varying parameters in smooth transition GARCH models. These non-linear models allow the dynamics of the conditional variance model to be influenced by the sign and size of past shocks. These factors can also be used as transition variables in the new smooth transition exponential smoothing approach. Parameters are estimated for the method by minimizing the sum of squared deviations between realized and forecast volatility. Using stock index data, the new method gives encouraging results when compared to fixed-parameter exponential smoothing and a variety of GARCH models.

Bradley and Jansen (2004) propose a model in which the dynamics that characterize stock returns are allowed to differ in periods following a large swing in stock returns—that is, a non-linear state-dependent model. Their approach allows them to test for the existence of non-linearities in returns, and to estimate the size of the shock that is required to cause the non-linear behavior.

4. Empirical issues

In this section we discuss various practical issues. In contrast to linear models, the design of non-linear models for actual data and the estimation of parameters are less straightforward.

How should we select a model? Should all the observations be used, or should models be estimated and/or evaluated against specific (dynamic) features of the historical record? Should we split the sample into in- and out-of-sample periods, and if so, where should the split occur?

Boero and Marrocu (2004) provide evidence related to some of these questions. They analyze the out-of-sample performance of SETAR models relative to a linear AR and a GARCH model using daily data for the Euro effective exchange rate. Their evaluation is conducted on point, interval and density forecasts, unconditionally, over the whole forecast period, and conditional on specific regimes. Their results show that the GARCH model is better able to capture the

---

4 There is a growing literature on nonparametric and semi-parametric forecasting methods and models. Traditionally this literature has focused on the conditional mean, but a number of recent papers have looked at nonparametric estimation of other aspects of conditional distributions, such as quartiles, intervals and density regions; see e.g. De Gooijer, Gannoun and Zerom (2002); Matzner-Löber, Gannoun and De Gooijer (1998); Samanta (1989), and the references therein. Developments in these areas look set to continue apace with the more parametric approaches considered in this issue.
distributional features of the series and to predict higher-order moments than the SETAR models. However, their results also indicate that the performance of the SETAR models improves significantly conditional on being in specific regimes. In a related study, Corradi and Swanson (2004) focus on various approaches to assessing predictive accuracy. One of the main conclusions of their study is that there are various easy to apply statistics that can be constructed using out-of-sample conditional-moment conditions, which are robust to the presence of dynamic misspecification. Because estimated models are approximations to the DGP and likely to be mis-specified in unknown ways, tests that are robust in this sense are obviously desirable. They provide an illustration of model selection via predictive ability testing involving the US money-income relation, and demonstrate the relevance of the various testing methods.

Next, which non-linear models should be entertained in any specific instance? This question can be answered by looking at the theoretical properties of the models, as well as the specific properties of the data under scrutiny. For example, Dahl and Hylleberg (2004) give a nice review of four non-linear models, namely, Hamilton’s flexible non-linear regression model (Hamilton, 2001), artificial neural networks and two versions of the projection pursuit regression model. The forecasting performance of these four approaches is compared for US industrial production and an unemployment rate series, in a ‘real-time forecasting’ exercise, whereby the specification of the model is chosen, and the parameters re-estimated at each step, as the forecast origin moves through the available sample. The paper provides some guidance for model selection, and evaluates the resulting forecasts using standard MSE-related criteria as well as direction-of-change tests. Interestingly, they find evidence that some of the flexible non-linear regression models perform well relative to the non-linear benchmark.

Further, how can we reliably estimate the model parameters? Can we address the problem whereby we only find local optima? How can we select good starting values in non-linear optimization? Can we design methods to test the non-linear models, based on in-sample estimation and in-sample data? How should we aggregate and analyze our data? Some of these questions are examined in van Dijk and Franses (2000), who consider daily, weekly and monthly data, and find distinct models for different temporal aggregation levels. But, what happens if we aggregate over the cross-section dimension? Marcellino (2004) argues that cross-section aggregation of countries, with constant weights over time, may produce ‘smoother’ series better suited for linear models, while aggregation with time-varying weights (and the presence of common shocks) is more likely to generate a role for non-linear modeling of the resultant series. Marcellino (2004) fits a variety of non-linear and time-varying models to aggregate EMU macroeconomic variables, and compares them with linear models. He assesses the quality of these models in a real-time forecasting framework. It is found that often non-linear models perform best.

Of course the bottom line is: Are there clear-cut examples where actual forecast improvements are delivered by non-linear models? Provision of such examples might serve to allay the fears held by many sceptics. Three nice recent examples of this sort are Clements and Galvão (2004); Sensier et al. (2004); Gençay and Selçuk (2004), all of whom show the relevance of non-linear models for forecasting. The latter paper suggests that the use of non-linearities is crucial for making sensible statements about the tail behavior of asset returns. In particular, Gençay and Selçuk (2004) investigate the relative performance of VaR models with the daily stock market returns of nine different emerging markets. In addition to well-known modeling approaches such as the variance-covariance method and historical simulation, they employ extreme value theory (EVT) to generate VaR estimates and provide the tail forecasts of daily returns at the 0.999 percentile along with 95% confidence intervals for stress testing purposes. The results indicate that EVT based VaR estimates are more accurate at higher quantiles. According to estimated Generalized Pareto Distribution parameters, certain moments of the return distributions do not exist in some countries. In addition, the daily return distributions have different moment properties in their right and left tails. Therefore, risk and reward are not equally likely in these economies. The other two papers consider macroeconomic variables. Sensier et al. (2004) examine the role of domestic and international variables for predicting classical business cycles regimes in four European countries, where
the regimes are classified as binary variables. One finding is that composite leading indicators and interest rates of Germany and the US have substantial predictive value. Clements and Galvão (2004) test whether there is non-linearity in the response of short and long-term interest rates to the spread. They assess the out-of-sample predictability of various models and find some evidence that non-linear models lead to more accurate short-horizon forecasts, especially of the spread. And, as mentioned, De Gooijer and Vidiella-i-Anguera (2004) report more marked gains to allowing for non-linearities.

5. Concluding remarks

In this paper we have summarized the state-of-the-art in forecast construction and evaluation for non-linear models, and in selection among alternative non-linear prediction models. We conclude that the day is still long off when simple, reliable and easy to use non-linear model specification, estimation and forecasting procedures will be readily available. Nevertheless, there are grounds for optimism. The papers in this issue suggest that careful application of existing techniques, and new models and tests, can result in significant advances in our understanding. Supposing that the world is inherently non-linear, then as computational capabilities increase, more complex models become amenable to analysis, allowing the possibility that future generations of models will significantly outperform linear models, especially if such models become truly multivariate.

Much remains to be done in the areas of specification, estimation, and testing, with important issues of non-differentiability, parameter-estimation error and data-mining remaining to be addressed. In addition, a further area for research (both empirical and theoretical) is the following. Suppose one has data with trend components, seasonality, non-linearity and outliers. How should one proceed? This is a complex issue, due to the possible inter-reactions between these elements, e.g. neglecting outliers may suggest non-linearity (see van Dijk, Franses & Lucas, 1999) and the trend can be intertwined with seasonality, as in periodic models of seasonality (see Franses & Paap, 2004 for an up-to-date survey). Much has been said about modeling each of these characteristics in isolation, but developing coherent strategies for such data remains an important task. It will be interesting to see to what extent developments in these areas give rise to tangible gains in terms of forecast performance—we remain hopeful that great strides will be made in the near future.

Acknowledgements

The authors wish to thank Valentina Corradi and Dick van Dijk for helpful conversations, and are grateful to Jan De Gooijer for reading and commenting on the manuscript. Swanson gratefully acknowledges financial support from Rutgers University in the form of a Research Council grant.

References


Hamilton, J. D. (1996). This is what happened to the oil price-

Hamilton, J. D. (2000). *What is an oil shock?* Mimeo. La Jolla, CA: Department of Economics, UCSD.


Tiao, G. C., & Tsay, R. S. (1994). Some advances in non-linear and


**Biographies:**

Michael P. CLEMENTS is a Reader in the Department of Economics at the University of Warwick. His research interests include time-series modeling and forecasting. He has published in a variety of international journals, has co-authored two books on forecasting, and co-edited (with David F. Hendry) of *A Companion to Economic Forecasting* 2002, Blackwells. He is also an editor of the *International Journal of Forecasting*.

Philip Hans FRANSES is Professor of Applied Econometrics and Professor of Marketing Research, both at the Erasmus University Rotterdam. He publishes on his research interests, which are applied econometrics, time series, forecasting, marketing research and empirical finance.

Norman SWANSON, a 1994 University of California, San Diego Ph.D. is currently Associate Professor at Rutgers University. His research interests include forecasting, financial- and macro-econometrics. He is currently an associate editor of the *Journal of Business and Economic Statistics*, the *International Journal of Forecasting*, and *Studies in Non-linear Dynamics and Econometrics*. He has recently been awarded a National Science Foundation research grant entitled the Award for Young Researchers, and is a member of various professional organizations, including the Econometric Society, the American Statistical Association, and the American Economic Association. Swanson has recent publications in *Journal of Development Economics*, *Journal of Econometrics*, *Review of Economics and Statistics*, *Journal of Business and Economic Statistics*, *Journal of the American Statistical Association*, *Journal of Time Series Analysis*, *Journal of Empirical Finance* and *International Journal of Forecasting*, among others. Further details about Swanson, including copies of all papers cited above, are available at his website: http://econweb.rutgers.edu/nswanson/.