A Multi-Step Forecast Density

Sebastiano Manzan[†] and Dawit Zerom[‡]

[†] CeNDEF, University of Amsterdam, the Netherlands

[‡] School of Business, University of Alberta, Canada

Abstract

This paper makes two contribution to the literature on density forecasts. First, we propose a novel bootstrap approach to estimate forecasting densities based on nonparametric techniques. The method is based on the Markov Bootstrap that is suitable to resample dependent data. The combination of nonparametric and bootstrap methods delivers density forecasts that are flexible in capturing markovian dependence (linear and nonlinear) occurring in any moment of the distribution. Second, we improve the testing approach to evaluate density forecasts by considering a set of tests for dynamical misspecification such as autocorrelation, heteroskedasticity and neglected nonlinearity. The approach is useful because rejections of the tests give insights into ways to improve the forecasting model. By Monte Carlo simulations we show that the proposed evaluation strategy has much higher power to detect misspecification of the density forecasts compared to previous analysis. The proposed nonparametric-bootstrap forecasting method exhibits the ability to capture correctly the dynamics of linear and nonlinear time series models. We also investigate the performance at higher orders and propose methods to deal with the "curse of dimensionality". Finally, we empirically investigate the relevance of the method in out-of-sample forecasting the density of 3 business cycles variables for the US: real GDP, the Coincident Indicator and Industrial Production. The results indicate that the method gives reliable density forecasts for all variables and performs better compared to parametric forecasting methods.

1 Introduction

Linear processes are often used to model and forecast economic time series. However, since the past two decades there has been a growing demand for models that can accommodate time series patterns that depart from the usual assumptions of linearity and Gaussianity. In fact, it is now extensively documented that economic time series exhibit significant nonlinear dependence. Hamilton (1989) provides evidence that US real GNP is well explained by a markov switching model in which the economy moves from a recessionary state with negative growth to an expansionary state with positive growth. This evidence suggests that nonlinear models are able to offer new insights in the dynamics of macroeconomic variables. Granger and Terasvirta (1993) reported successful application of these models to a variety of economic time series. Extended surveys of nonlinear models are also provided by Tong (1990), Krolzig (1997) and Terasvirta (1998). However, the significant in-sample evidence has not translated into improved forecastability out-of-sample. de Gooijer and Kumar (1992) and more recently Clements *et al.* (2004) review a large body of literature showing the failure of nonlinear models in outperforming linear ones in out-of-sample prediction. An explanation is offered by Pesaran and Potter (1997). They suggest that nonlinearities might be more effectively evaluated when forecasts of higher moments (such as the conditional variance) are considered. This indicates that a more appropriate comparison of linear and nonlinear models should examine the density forecasts rather than point forecasts.

Another argument in favor of making use of density forecasts is provided by Granger and Pesaran (2000a, 2000b). They argue that focusing on point forecasts is justified when the decision problem faced by agents is linear in constraints and quadratic in loss function. In practice, economists are interested in evaluating decisions that involve events such as recessions or that inflation will be in a certain interval. This has also contributed to shift the attention toward producing density rather than point forecasts. The advantage of density forecasts is that they provide an estimate of the future probability distribution of a financial or economic variable, conditional on the information available at the time the forecast is made. In that sense, a density forecast represents a complete characterization of the uncertainty associated with the prediction. The practical relevance is also demonstrated by the decision of the Bank of England and the Royal Bank of Sweden to report the uncertainty around their point forecast of inflation and GDP.

The surge of interest for density forecasts required the development of suitable statistical tools for their evaluation and comparison. Corradi and Swanson (2004) provide a survey of the work in the field. The first method was proposed by Diebold *et al.* (1998). Their method is very convenient because it

transform the problem of evaluating the conditional density into the problem of testing the properties of the Probability Integral Transform (PIT). Hong *et al.* (2004) and Hong and Li (2005) compare different models for the spot interest rate based on the evaluation of the density forecasts. In a recent paper, Clements *et al.* (2003) investigate the power of point and density forecasts tests to distinguish between linear and nonlinear forecasting models when the true model is in fact nonlinear. Their results cast doubts on the ability of the evaluation tests to correctly identify the misspecification of the linear density forecasts at the typical sample size of macroeconomic time series. This suggests that the forecasting failure of nonlinear models could be attributed to the methodology used to evaluate predictability rather than the inability of nonlinear models to capture genuine dependence in the data.

In this paper we make two contributions to the existing literature. First, we propose a nonparametric bootstrap method for generating density forecasts. The method is data-driven and is preferable when the analyst is uncomfortable with prior assumptions regarding the form of the dependence (e.g., linear or nonlinear) and/or the distribution of the error term (e.g., Gaussian). We adopt a novel approach to estimate the conditional density that combines the flexibility of nonparametric methods with bootstrap techniques. The method is based on a resampling technique for dependent data called Markov Bootstrap. An early reference is Rajarshi (1990) and more recent work in the econometrics literature are Paparoditis and Politis (2001, 2002) and Horowitz (2003). The method is able to account for any type of (markovian) dependence that occurs in the conditional distribution, such as in the conditional mean, variance and/or skewness. The assumption of markov dependence is not restrictive as it encompasses a wide range of relevant structures implied by various commonly used linear and nonlinear models (e.g., AR and SETAR models). Other nonparametric methods to estimate the forecasting density were proposed by Hyndman and Yao (2002) and de Gooijer and Zerom (2003). Unlike these direct estimators, our method is an empirical one where density forecasts are produced as outcomes of the Markov bootstrap procedure.

The second contribution of the paper is to elaborate an appropriate testing strategy to evaluate density forecasts. As mentioned earlier, Clements *et al.* (2003) report the low power of density evaluation methods to detect the neglected nonlinearity. We augment the evaluation stage with suitable tests for serial independence, ARCH and linearity. This addresses the problem of the low power of existing tests to capture the misspecification of the dynamics of the forecasting densities.

The paper is structured as follows. In Section (2) we briefly review parametric bootstrap methods to estimate the forecasting density and then introduce the nonparametric bootstrap method. We discuss also the issue of the bandwidth choice. In Section (3) we present the testing approach. In Section (4) we present the results of the Monte Carlo experiment. We investigate the power of the

density forecasts tests and the performance of the Markov Bootstrap for various linear and nonlinear time series models. In Section (5) we apply the method to the quarterly growth rate of US GDP and to monthly growth rate of the Conference Board Coincident Index and US Industrial Production. Finally, Section (6) concludes.

2 Estimation of the τ-step forecast density

Let $Y_1, \cdots, Y_T$ denote realizations of a time series process. We consider the problem of forecasting future values $Y_{T+1}, \cdots, Y_{T+\tau}$ based on the observed data $Y_T, Y_{T-1}, \cdots$. In particular, our goal is to estimate the out-of-sample τ-step forecast density for $\tau \geq 1$. The τ-step forecast density is the conditional density of $Y_{T+\tau}$ given $Y_T, Y_{T-1}, \cdots$, i.e. $f_{T+\tau}(y|\cdot)$ where $y \in \mathbb{R}$. In tackling this problem, by far the most common approach is to model the dependence of future observations on the past and present via a parametric class of models.

Here, we introduce a nonparametric approach in which the estimation of the forecast density does not require a priori specification of a model. To that purpose, we assume that the time series Y_t is the outcome of a p-th order Markov process,

$$\mathbf{P}(Y_{t+1} \leq y_{t+1}|Y_t = y_t, Y_{t-1} = y_{t-1}, \ldots) = \mathbf{P}(Y_{t+1} \leq y_{t+1}|Y_t = y_t, Y_{t-1} = y_{t-1}, \ldots, Y_{t-p} = y_{t-p})$$

almost surely for $y_{t+1}, y_t, y_{t-1}, \ldots$ and some finite integer $p \geq 1$. The assumption of Markov dependence is satisfied by a large class of linear and nonlinear models that are of interest in time series analysis and forecasting. One frequently used special case of a Markov process is the conditionally heteroskedastic autoregressive model of the following form

$$Y_{t+1} = \mu(X_t, \theta) + \sigma(X_t, \beta)\epsilon_{t+1} \tag{1}$$

where $X_t = (Y_t, \ldots, Y_{t-p+1})'$ is a p-dimensional vector, $\mu(\cdot)$ is the conditional mean of the process, $\sigma(\cdot)$ denotes the conditional variance, and ϵ_{t+1} is an $i.i.d$ disturbance term with zero mean. The vectors θ and β denote the parameters in the conditional mean and the conditional variance of the process, respectively. Model (1) includes several familiar time series models such as linear AR, ARCH, and SETAR.

To motivate our method for estimating the forecast density, we first briefly review how model-based τ-step forecast densities can be estimated via bootstrapping. An extensive account is given in Clements and Smith (1997). The method to be introduced in section (2.1) can be interpreted as a nonparametric generalization of the model-based bootstrap approach. The motivation to use the

bootstrap in a parametric framework is twofold. First, the researcher might find restrictive to make assumptions about a parametric distribution for the innovation ϵ_{t+1}. Second, for many nonlinear models of interest there are no analytical expressions for the expectation τ-step ahead. Under the model specification in (1) and conditional on current time period, say T, the bootstrap realization one-step ahead is given by

$$Y^*_{T+1,b} = \mu(X_T, \hat{\theta}) + \sigma(X_T, \hat{\beta})\hat{\epsilon}^*_{t+1,b}$$

where $\hat{\epsilon}^*_{t+1,b}$ are the *i.i.d.* resampled (bootstrap) residuals, and $\hat{\theta}$ and $\hat{\phi}$ are consistent estimators of θ and β, respectively. The 1-step ahead forecast density, $f_{T+1}(\cdot|X_T)$, at time T is then given by the empirical density function of the bootstrap realizations, $y^*_{T+1,b}$, $b = 1, \ldots, B$ where B is the desired number of replications. For $\tau \geq 2$, the forecasting density can be obtained by applying an iterative scheme. The 2-step ahead conditioning vector is updated to $X^*_{T+1,b} = (Y^*_{T+1,b}, Y_T, \cdots, Y_{T-p+2})$ and the forecast density, $f_{T+2}(\cdot|X_T)$, is again the empirical density of the bootstrap realizations

$$Y^*_{T+2,b} = \mu(X^*_{T+1,b}, \hat{\theta}) + \sigma(X^*_{T+1,b}, \hat{\beta})\hat{\epsilon}^*_{t+2,b},$$

for $b = 1, \ldots, B$. The forecast τ-step ahead is obtained by iterating the bootstrap procedure.

2.1 Markov Forecast Density

The model-based approach of forecast density estimation is a residual-based procedure in the sense that it begins by estimating θ and β and subsequently uses *i.i.d.* resampling of the fitted residuals to form bootstrap realizations. On the other hand, the Markov Forecast Density (hereinafter MFD) estimator does not attempt to reduce the problem to *i.i.d.* residuals. Instead, an appropriate resampling procedure is applied directly to the realizations Y_t. In comparison to the model-based procedure, the MFD offers at least two advantages. First, the nonparametric nature of the MFD avoids making assumptions of a parametric specification that may be inappropriate. Second, MFD is valid for a class of time series models that is wider than (1). The method is able to account for dynamics that occur in any moment of the conditional distribution, such as the conditional mean and variance but also, for example, the conditional skewness. In this sense it is a more general method compared to the model assumed in Equation (1). The MFD procedure is outlined below.

Assume that the Markov order p of the time series realization Y_t is known. Cheng and Tong (1992) and Diks and Manzan (2002) propose nonparametric methods to select p. Suppose the forecast origin is at time $t = T$ and define the corresponding conditioning vector by $X_T = (Y_T, Y_{T-1}, \ldots, Y_{T-p+1})'$. The object of interest is to estimate the out-of-sample τ-step forecast density $f_{T+\tau}(\cdot|X_T)$ using the available data $Y_1, \ldots, Y_T$. For $t = p, \ldots, T-1$, define the vectors $X_t = (Y_t, Y_{t-1}, \ldots, Y_{t-p+1})'$. The

strategy is to assign probability weights to each vector $X_p, \cdots, X_{T-1}$, and using those probabilities to resample from their successors. The values of these probabilities will depend on the "closeness" of the vectors X_t to the conditioning vector X_T. Those states that are "close" to X_T receive larger probability weights compared to those that are further away. Thus, by suitably choosing these probability values, we maintain the Markov dependence in the data. For the computation of the probability distribution, a nonparametric (kernel smoothing) approach is used. The algorithm is an extension of a recently popularized local bootstrap approach by Paparoditis and Politis (2001, 2002) to the context of out-of-sample forecast density estimation.

The algorithm is as follows:

Step 1 For $t = p, \ldots, T-1$, define probability weights

$$P(J = t) = K\Big(\frac{X_T - X_t}{h_T}\Big) / \sum_{m=p}^{T-1} K\Big(\frac{X_T - X_m}{h_T}\Big). \tag{2}$$

where $h_T > 0$ is a bandwidth parameter and $K(\cdot)$ is a kernel function. $K(\cdot)$ should be a symmetric probability density. Using the probability mass function above, draw with replacement from the successors of X_t, i.e., $Y^*_{T+1} = Y_{J+1}$.

Step 2 For $\tau \geq 2$:

Step 2.1 Define the new conditioning feature vector at time $t = T+1$ as $X^*_{T+1} = (Y^*_{T+1}, Y_T, \ldots, Y_{T-p+2})$ and calculate the new probability weights

$$P(J = t) = K\Big(\frac{X^*_{T+1} - X_t}{h_T}\Big) / \sum_{m=p}^{T-1} K\Big(\frac{X^*_{T+1} - X_m}{h_T}\Big).$$

Using the updated probability mass function above, draw with replacement from the successors of X_t, i.e., $Y^*_{T+2} = Y_{J+1}$.

Step 2.2 Continue from Step (2.1) until forecasts for period $T+\tau$ are obtained using a conditioning vector $X^*_{T+\tau-1} = (Y^*_{T+\tau-1}, Y^*_{T+\tau-2}, \ldots, Y^*_{T+1}, \ldots, Y_{T-p+\tau})$.

Step 3 Repeat Step (1) (and (2) if $\tau \geq 2$) B times.

The above procedure creates a distribution of values for each τ given by $Y^{*,1}_{T+\tau}, Y^{*,2}_{T+\tau}, \ldots, Y^{*,B}_{T+\tau}$. The τ-step MFD estimator, $\hat{f}_{T+\tau}(\cdot|X_T)$, is defined as the kernel density estimate of these B-observations. The MFD estimator for $\tau = 1$ can be written as

$$\hat{f}_{T+1}(y|X_T) = \sum_{j=p}^{T-1} W_{a_T}\Big(\frac{y - Y_{j+1}}{a_T}\Big) P(Y^*_{T+1} = Y_{j+1}|X_t = X_T) \tag{3}$$

where $y \in \mathbb{R}$, $W_{a_T}(\cdot) = 1/a_T W(\cdot)$ in which $W(\cdot)$ is a kernel function, and $a_T > 0$ is a bandwidth parameter. The asymptotic validity of this estimator is discussed in Section (2.2). In Section (4), we empirically evaluate the finite-sample performance of MFD using a battery of specification tests.

The above algorithm for calculating the MFD estimator may seem involved. But, essentially, its implementation only requires two simple steps. Take the case of $\tau = 1$. In the first step, we calculate the weights as in equation (2) and use these probability mass to resample the data with replacement, i.e. $Y^{*,1}_{T+1}, Y^{*,2}_{T+1}, \ldots, Y^{*,B}_{T+1}$. In the second step, we compute the kernel density estimate based on the resampled data. Most statistical software packages have routines to perform the second step.

2.2 Theoretical justification for the MFD estimator

We begin with $\tau = 1$. Using equation (2), i.e., $P(Y^*_{T+1} = Y_{j+1}|X_t = X_T) = P(\cdot)$, we can write equation (3) as

$$\hat{f}_{T+1}(y|X_T) = \sum_{t=p}^{T-1} W_{b_T}\Big(\frac{y - Y_{t+1}}{a_T}\Big) K\Big(\frac{X_T - X_t}{h_T}\Big) \Big/ \sum_{t=p}^{T-1} K\Big(\frac{X_T - X_t}{h_T}\Big). \tag{4}$$

We can see from the above equation that our 1-step MFD estimator is essentially the classical Nadaraya-Watson estimator of the conditional density suitably adapted to forecasting. Now, let (X_t, Z_t), where $Z_t = Y_{t+1}$, be a sequence of $\mathbb{R}^p \times \mathbb{R}$-valued strictly stationary process. The consistency of the conditional density estimator was proved by Hyndman *et al.* (1996) for the case of independent (X_t, Z_t), and by Fan *et al.* (1996) and de Gooijer and Zerom (2003) in the case of dependence. Under certain mixing conditions on the process (X_t, Z_t) and some technical regularity assumptions, $\hat{f}_{T+1}(y|X_T)$ is a consistent estimator of $f_{T+1}(y|X_T)$, i.e., as $T \to \infty$,

$$\hat{f}_{T+1}(\cdot|X_T) \xrightarrow{\mathcal{P}} f_{T+1}(\cdot|X_T). \tag{5}$$

As outlined in the algorithm, when $\tau \geq 2$, the MFD estimator is defined by repeating one-step ahead predictions τ times, treating the bootstrap value from the last round as the true value. In this way, the τ-step MFD can be viewed as a one-step plug-in estimator. The asymptotic consistency result for the 1-step MFD also holds for τ-step MFD. We only need to show that replacing actual values by bootstrap replicates is valid. For example, for $\tau = 2$, it suffices to show the validity of replacing Y_{T+1} by the bootstrap counterpart Y^*_{T+1}. Let $\mathcal{C}$ denotes a fixed compact subset of $\mathbb{R}^p$ on which the marginal density of x is lower bounded by some positive constant. Under a set of regularity conditions, Paparoditis and Politis (2001, 2002) show that the one-step transition distribution function $F^*(y|x)$ that governs the law of the Markov bootstrap process satisfies the following uniform convergence property (see Theorem 3 in Paparoditis and Politis (2002)): $\sup_{y\in R} \sup_{x\in\mathcal{C}} |F^*(y|x) - F(y|x)| \to 0 \quad a.s$ where $F(y|x)$ is the

true one-step transition distribution function. Now, replacing the conditioning vector x by X_T, it is easy to see that

$$\sup_{y\in\mathbb{R}} |F^*(y|X_T) - F(y|X_T)|\mathbf{1}(X_T \in \mathcal{C}) \leq \sup_{y\in R}\sup_{x\in\mathcal{C}} |F^*(y|x) - F(y|x)|$$

where $\mathbf{1}(A)$ denotes the indicator function for set A. Therefore,

$$\sup_{y\in\mathbb{R}} |F^*(y|X_T) - F(y|X_T)| \to 0 \quad a.s.$$

Because $F^*(y|X_T)$ is the law that generates Y^*_{T+1} and $F(y|X_T)$ is the law that generates Y_{T+1}, we can replace Y_{T+1} by Y^*_{T+1}. For $\tau > 3$, the same argument holds by induction.

2.3 Choice of bandwidth

As is the case for all kernel-based nonparametric methods, the asymptotic validity of our multi-step forecast density estimator requires the bandwidth parameters $h_T \to 0$ and $a_T \to 0$ as $T \to \infty$. However, in practice the sample size T is fixed. Thus, some decisions have to be made before calculating the forecast density. Note that, unlike other nonparametric conditional density estimators, our estimator does not require both a_T and h_T to be chosen simultaneously. By construction, the estimator is implemented in two stages. Furthermore, detailed simulation experiments suggest that our estimator is not sensitive to the choice of a_T as long as $a_T \sim T^{-1/5}$. On the other hand, the choice of h_T is critical to the quality of the forecast density estimator.

For fixed T, when $h_T \to 0$, the τ-step forecast density will tend to accurately capture the dependence structure or dynamical properties of the data. The problem is that the forecast density becomes excessively peaked compared to the true density; see the Monte Carlo simulation experiment Section for more on this situation. On the other hand, when $h_T \to \infty$, the τ-step forecast density does not reflect the dependence structure of the data. The latter case represents a situation where the data are in fact independent. Notice that when $h_T \to \infty$, the probability weight $P(J = s) \to 1/(T-p)$ suggesting the information contained in X_T is irrelevant. Therefore, to obtain a forecast density that is well behaved while accurately mimicking the dependent characteristics of the data, h_T should lie between the above two extremes.

The probabilities $P(\cdot)$ used to resample the data are just kernel estimates scaled by kernel densities. We suggest a two-step bandwidth selection procedure. We first estimate a pilot density estimate using h_T that is suited for *i.i.d.* data, i.e. $h_T(fixed) = \hat{\sigma}T^{-\frac{1}{p+4}}$, where $\hat{\sigma}$ is the standard deviation of the sample (i.e. a scale parameter for the data). As is well known, this fixed bandwidth is not adaptive to the data configuration of X_t. In particular, it gives inaccurate estimates in those regions of the data

where observations are scarce. The problem of data scarcity for the fixed bandwidth becomes even more serious when the data are time series. Dependence has a tendency to accentuate the problem of data sparsity. Thus, we use the pilot density estimate to adjust the fixed bandwidth in such a way that areas of high density use a small bandwidth and areas of low density use a larger bandwidth. Following Silverman (1986), we define local bandwidth factor, λ_t, by $\lambda_t = \left\{ \frac{\widehat{f}(X_t)}{g} \right\}^{-\alpha}$ where g is the geometric mean of $\widehat{f}(X_t)$, i.e. $\log g = \frac{1}{T} \sum_{t=1}^{T} \log \widehat{f}(X_t)$, and α $(0 \leq \alpha \leq 1)$ denotes the sensitivity parameter that regulates the amount of weight that is attributed to the observations in the low density regions. We fix $\alpha = 0.5$. Using the local bandwidth factor λ_t, an adaptive bandwidth is defined as $h_T(adaptive) = \lambda_t h_T(fixed)$. In Section (4), we compare the performances in finite samples of both fixed and adaptive bandwidths.

3 Forecasting Densities Evaluation

In the previous Section we proposed a nonparametric method to forecast the density of a time series. The next step consists of evaluating the goodness of the densities to explain the conditional distribution of the data. In regression analysis the model is evaluated by testing the residuals against various forms of misspecification. Rejections of the residuals tests indicate that the model is inappropriate and that improvements are required. In the context of density forecasts a similar approach is proposed by Diebold *et al.* (1998) based on the Probability Integral Transform (PIT). The method is convenient because it transforms the problem of testing the goodness of the forecasting densities into the problem of testing the properties of the PIT random variable. Under the hypothesis of correct specification, the PIT is *i.i.d.* uniformly distributed in the interval [0,1]. By testing the properties of the PIT it is possible to evaluate the goodness of the conditional density in explaining the distribution and the dependence in the time series. Tests used for the residuals of a regression can be extended to testing the PIT. Denote by $f(Y_{t+\tau}|X_t)$ the true conditional density of the time series and by $\hat{f}(Y_{t+\tau}|X_t)$ the forecasting density. We assume that the forecasting exercise starts at time T and the aim is to predict τ step ahead the next P realizations of the process. We first consider the case of testing the density forecasts for $\tau = 1$ and we will discuss later in the section the approach to test when $\tau \geq 2$. The PIT denoted by z_s, for $s = 1, \cdots, P$, is defined as

$$z_s = \int_{-\infty}^{y_{T+s}} \hat{f}_{T+s-1}(u)du \tag{6}$$

where $\hat{f}_{T-s+1}(\cdot)$ indicates the forecasting density at time $T + s - 1$. Under the assumption that $f_{T-s+1}(\cdot) = \hat{f}_{T-s+1}(\cdot)$, z_s is *i.i.d.* uniformly distributed in the interval [0,1]. Diebold *et al.* (1998)

propose to evaluate the goodness of the forecasting density by testing the uniformity of the PIT using the Kolmogorov-Smirnov (*KS*) test. However, this test evaluates only the distributional part of the assumption and not the independence of the PIT. This issue explains the results of Clements *et al.* (2003). They investigate the power of the uniformity test to distinguish between a linear and nonlinear forecasting density when the DGP is actually nonlinear. They show that it has negligible power to indicate the misspecification of the linear density at the sample sizes typically available for macroeconomic time series. Recently, Hong *et al.* (2004) and Hong and Li (2005) use a test of the joint hypothesis of uniformity and *i.i.d.* based on the generalized spectrum. They also consider test statistics of the hypothesis of dependence occurring for powers of z_s. This is a relevant information because it suggests directions in which the forecasting density could be improved to correctly account for the dynamics of the series. In this paper, we propose a similar approach to evaluate the independence assumption of the PIT. We test for specific directions of dynamic misspecification, such as autocorrelation, ARCH and nonlinearity in the PIT. To achieve this goal we adopt the following testing strategy. We assume that the PIT z_s follows the process

$$(z_s - \bar{z}) = \alpha_1(z_{s-1} - \bar{z}) + \cdots + \alpha_q(z_{s-q} - \bar{z}) + \epsilon_s \tag{7}$$

and test the hypothesis that all the α_i's (for $i = 1, \cdots, q$) are equal to zero. This can be carried out using an LM-type test with statistic equal to P times the R^2 of Equation (7) that is distributed as a $\chi^2(q)$. Rejection of the null hypothesis suggests that there is linear dependence unaccounted by the forecasting model. We will denote the test for serial correlation as SC in the following sections. Another alternative of interest is that the density forecasts do not correctly account for ARCH structure. To test against this hypothesis we perform an ARCH LM test. We regress the squared residuals of Equation (7) on r lags

$$\epsilon_s^2 = \beta_1\epsilon_{s-1}^2 + \cdots + \beta_r\epsilon_{s-r}^2 + \eta_s \tag{8}$$

and test the null hypothesis that the β_j (for $j = 1, \cdots, r$) are jointly equal to zero. The test statistic is P times the R^2 of Equation (8) and it is distributed as a $\chi^2(r)$. We denote this test as $ARCH$ in the simulation and empirical part. In addition, we are also interested in evaluating the ability of the forecasting model to account for the nonlinearity of the underlying process. We use the *V23* test proposed by Terasvirta *et al.* (1993) to test the hypothesis of neglected nonlinearity in the conditional mean of the PIT. We interpret the rejection of the hypothesis as evidence that the forecasting density does not account for nonlinear dynamics of the time series under investigation. We denote this test as $V23$.

We described the testing approach for one-step ahead density forecasts. When the interest is a τ larger than 1, it should be taken into account the $(\tau - 1)$ dependence of the forecasts. Diebold *et al.* (1998) propose an easy approach to solve this problem. They suggest to consider the sub-series $(z_1, z_{1+\tau}, z_{1+2\tau}, \cdots)$, $(z_2, z_{2+\tau}, z_{2+2\tau}, \cdots)$ and $(z_\tau, z_{2\tau}, z_{3\tau}, \cdots)$ that should be *i.i.d.*. The same battery of tests described above can be applied to the sub-series of the PIT using a significance level equal to $\frac{\alpha}{\tau}$, where α is the size of the test and τ the forecasting step. The null hypothesis is rejected if any of the τ tests rejects.

4 Monte Carlo Results

We examine by simulation the ability of the MFD to correctly account for the dynamics of some linear and nonlinear time series models. We proceed as follows. We simulate a time series and predict recursively τ-step ahead from observation T to $T - \tau$. We then test the P out-of-sample forecasts with the battery of misspecification tests introduced in the previous section. In all simulations we limit the analysis to $\tau = 1$. We consider three (in-)sample sizes T equal to 300, 600 and 900 and values for the prediction period P equal to 100 and 300. This means that for T = 300 and P = 100 we generate a series of length 400 and forecast the last 100 observations. We fix the number of bootstrap replications to 1000 and the number of simulations to 2000. We consider first and second order markovian processes. The higher dimensional case allows the evaluation of the incidence of the curse of dimensionality and the benefit (if any) of adopting the adaptive bandwidth.

4.1 Markov processes of order 1

We consider the following time series models:

AR(1): $y_t = \phi y_{t-1} + \epsilon_t$

ARCH(1): $y_t = \sigma_t \epsilon_t$ and $\sigma_t = (1 - \alpha) + \alpha y_{t-1}^2$

SETAR(1): $y_t = [-1.25 - 0.7y_{t-1} + \sigma_1 \epsilon_t] I(y_{t-1} \leq r) + [0.3y_{t-1} + \sigma_2 \epsilon_t] I(y_{t-1} > r)$

where ϵ_t is standard normally distributed and ϕ, α and (σ_1, σ_2, r) are parameters. The first model is a linear homoskedastic AR(1) process. The second model is an ARCH(1) process and the third represents a SETAR specification. For the last model we fix the parameters in the regimes and consider two situations, one in which the regimes are homoskedastic ($\sigma_1 = \sigma_2$) and the other in which they are heteroskedastic ($\sigma_1 \neq \sigma_2$). Clements *et al.* (2003) show that the *KS* uniformity test has very low power to detect the dynamical misspecification of linear forecasting densities when in fact

the underlying process is of the SETAR type above. In the previous section we argue that a more effective approach consists of investigating the dynamical properties of the PIT that could indicate the neglected dependence in the forecasting density. We perform a power analysis of the testing strategy and we later discuss the performance of MFD in forecasting the densities of the processes considered.

4.2 Power of the Tests

To investigate the power of the tests we draw with replacement past observations of the simulated series and assume that the forecasting density is given by the empirical distribution function of the bootstrap replications. In this way we destroy the serial dependence in the time series although maintaining the unconditional shape of the distribution.

Table (1) shows the frequency of rejections of the four tests at 5% significance level for the simulated processes. We consider both a forecasting period of 100 and 300 periods. The results show that the best performance of the *KS* test for uniformity occurs for the AR(1) model where it rejects in 22% of the simulations. For the other processes the power is lower. This confirms the result of Clements *et al.* (2003) that the *KS* test of the PIT is hardly informative about the dynamic misspecification of the forecasting densities. Instead, the *SC* test for first order autocorrelation of the *PIT* correctly detects that the density forecasts for the AR(1) process are misspecified. The rejection rate is close to 100% already for the 100 out-of-sample forecast period. As expected, for the ARCH(1) process the *SC* test rejects approximately 6.3% of the times. The SETAR specification has some linear dependence in addition to the nonlinear one. This is captured by the *SC* test that rejects (P=300) in 98% and 56% of the simulation for the homoskedastic and heteroskedastic cases, respectively. The *HET* test applied to the PIT of the ARCH process rejects in 32% of the simulations for P = 100 and in 83% of the cases in the longer sample. Although power increases slowly with sample size, there is an indication that the test correctly points to the lack of ARCH dynamics in the forecasting model. For the AR(1) and homoskedastic SETAR the rejection frequencies are close to the 5% nominal level. Some ARCH dependence is detected when the heteroskedastic SETAR is considered. The *HET* test rejects in 15% of the simulations for P=300. It is probably due to the change in the volatility level of the time series due to the switching between the two regimes. This suggests that a spurious ARCH effect can arise when nonlinear models are considered. Finally, the *V23* linearity test correctly suggests that the first two simulated models do not show evidence of neglected nonlinear structure in the conditional mean. Instead, the test has very high power against SETAR dynamics: in the homoskedastic case the rejections increase from 60% to 97% for the largest sample, and from 74% to 99% in the heteroskedastic case.

Table (1) here

Overall, the simulation results confirm previous evidence that testing the uniformity of the PIT does not have high power to detect the dynamic misspecification of the forecasting densities. In addition, it does not indicate a direction toward which improvements of the forecasting model can be made. However, the testing strategy proposed represents a practical way to investigate the directions of dynamical misspecification of the forecasting model. In particular, the *V23* test is a powerful approach when the analysts suspects the presence of nonlinearities in the data.

4.3 Performance of the MFD

The previous results suggest that for P=300 the tests have high power to detect the dynamic misspecification of the forecasting densities. In order to have a more compact presentation of the results we report only the rejection frequencies for the longest forecasting period. We implement the MFD with the fixed and adaptive bandwidths described in Section (2.3). We use a gaussian kernel and consider three bandwidths equal to 0.5, 0.75 and 1 time $h_T(fixed)$ and $h_T(adaptive)$ in Section (2.3). The objective of this exercise is to investigate the effect of the bandwidth on the performance of the proposed method.

Tables (2) to (5) report the frequencies of rejection of the null hypotheses of the tests for the MFD and for the linear parametric bootstrap as assumed in Equation (1) where $\mu(\cdot)$ is assumed linear and $\sigma(\cdot)$ is constant. We consider the performance of the linear forecasting densities as a benchmark for comparing the MFD.

Table (2) shows the frequencies of rejection (at 5% significance level) for the AR(1) process with coefficient equal to 0.5. In this case the linear parametric bootstrap is the appropriate forecasting method and the tests do not indicate misspecification. The MFD implemented with the fixed (rule-of-thumb) bandwidth captures correctly the dynamics in the simulated series for $c = 0.5$ but overrejects for $c = 1$: for T = 300 the rejection frequency increases from 5.6% to 9.3% for the largest value. However, increasing T to 600 and 900 observations contributes to correct the problem. The opposite situation occurs for the *HET* test when T = 300: increasing the bandwidth from $c = 0.5$ to 1 decreases the rejections from 14% to 6.9%. Also in this case larger samples contribute to drive the rejection frequency toward the nominal value. It is clear that the choice of the value of the bandwidth is a very relevant issue for the performance of the method. A too small bandwidth accounts correctly for the dynamics in the time series at the cost of a spuriously volatile PIT. This is due to the fact that the forecasting density is excessively peaked around the realization compared to the true density and creates periods

of small and large values of the PIT. This effect is picked up by the *HET* test. A too large bandwidth achieves the opposite effect. The *KS* and *V23* tests have frequencies very close to 5% in all cases. This evidence suggests that for T = 300 and $c = 0.75$ the MFD delivers quite good results although increasing the sample size improves significantly the performance. The only remarkable difference between the fixed and adaptive bandwidths concerns the *HET* test. The adaptive has lower rejection frequencies for all c considered: in the case of T = 600 and c equal to 0.5 the rejections decrease from 11% to 7.3% when using the adaptive.

Table (2) here

Summarizing the results for the AR(1), the MFD accounts reasonably well for the dynamics of the underlying model. Already for $T = 300$ the results are satisfactory, in particular when the adaptive bandwidth is considered. Undersmoothing using a bandwidth of around 0.75 times the standard value achieves the best results.

The results for the ARCH(1) model with $\alpha = 0.3$ are in Table (3). In this case, the linear forecasting densities are misspecified and this is clearly detected by the *HET* test that rejects in 83% of the simulations. Concerning the MFD, for all sample sizes and bandwidths the *KS*, *SC* and *V23* tests do not deviate significantly from the nominal rejection level. However, there are significant over-rejections for the *HET* test. For the fixed bandwidth and T = 300, the rejection frequency is 23% for $c = 0.5$ and 17% for the largest value. These values become smaller when the sample is increased to T = 900 but they are still significantly larger than the nominal level. The adaptive bandwidth performs better compared to the fixed one but it is still not satisfactory: it rejects in 17% of the simulations for T=300 and c=0.5 and in 14% for c=1. When the sample is increased to T = 900 the rejections range from 11% to 7.8% for c equal to 0.5 and 1, respectively. Although we only considered bandwidths ranging from 0.5 to 1 the standard values, it is clear that for this model there is a tendency to benefit from oversmoothing. Considering values of c larger than 1 would probably achieves better results.

Table (3) here

Table (4) shows the rejection frequencies for the homoskedastic SETAR model with $\sigma_1 = \sigma_2 = 1$ and $r = -0.20$. The *V23* test applied to the linear forecasting densities indicates the neglected nonlinearity with a rejection frequency of 94%. The *SC* and *HET* tests also overreject slightly. The *KS* test shows very low power to detect the misspecification of the conditional mean of the model. The *V23* test for the MFD clearly indicates that the method is correctly accounting for the nonlinear dependence in the data already for T = 300. This is a robust result across types and values of the

bandwidth. The *KS* and *SC* tests do not suggest misspecification of the forecasting densities. However, the spurious ARCH effect is present also for this model at the smallest value of the bandwidth. For c equal to 0.5 the rejections for the fixed bandwidth are 17% when T = 300 and 9.1% for T = 900. Instead, for c=1 the rejection frequencies are much closer to the nominal level: 7.6%, 7.3% and 6.5% as the sample size is increased from 300 to 900. The adaptive bandwidth performs slightly better compared to the fixed one. The MFD appears to be quite successful in modeling the nonlinearity of the homoskedastic SETAR model. The choice of a bandwidth close to the standard value is able to model correctly the dynamics and avoid the danger of the spurious ARCH effect.

Table (4) here

This simulation shows the main contributions of the paper. On the one hand, we propose a test that has very high power to detect the misspecification of the linear forecasting densities when the true density is in fact nonlinear. The second innovative aspect is that the proposed Markov Predictive Densities correctly account for the nonlinearity in the SETAR model when measured by the *V23* test. In this respect we offer a new method and evaluation strategy compared to the previously mentioned results of Clements *et al.* (2003).

The final model that we consider is a SETAR specification where the variances in the regimes are different. We assume that $\sigma_1 = 1$, $\sigma_2 = 2$ and $r = -0.10$. The *V23* test for the linear forecasting densities has almost unit power against the alternative. The misspecification of the parametric bootstrap shows up also in the overrejections for the *SC* and *HET* tests. The performance of the MFD depends on the value of the bandwidth: the nonlinearity is correctly accounted when c is equal to 0.5 and slightly deteriorates at the largest value (considering the case of T = 300). On the other hand, undersmoothing give rise to the spurious evidence of ARCH structure that declines when the bandwidth is large. The mid-value of c equal to 0.75 seems to balance between the two effects, in particular when the adaptive smoothing is considered. For this value of the adaptive bandwidth and T = 300 the *V23* and *HET* are slightly larger than the nominal level. However, for larger samples they seem to converge toward the correct rejection rate.

Table (5) here

4.4 Performance of the MFD at higher orders

The previous results involved Markovian processes of order 1. It is also interesting to investigate the performance of the MFD for higher orders and the role played by the adaptive bandwidth. Here we consider two variations of the homoskedastic SETAR model presented earlier:

SETAR(2): $y_t = [-1.25 - 0.7y_{t-2}]I(y_{t-2} \leq r) + 0.3y_{t-2}I(y_{t-2} > r) + \sigma\epsilon_t$

SETAR(1-2): $y_t = [-1.25 - 0.7y_{t-1}]I(y_{t-2} \leq r) + 0.3y_{t-2}I(y_{t-2} > r) + \sigma\epsilon_t$

where in SETAR(2) the dependence occurs only in the second lag whereas in SETAR(1-2) there is dependence both in the first and second lags. For generating the forecasting densities (linear and markovian) and for testing we assume that the order is equal to 2 and assume that $\sigma = 1$ and $r = -0.10$. In Table (6) and (7) we also report the rejection frequency for bootstrap under the null of independent data (indicated as IND). This let us evaluate the power of the tests in the higher order case.

Table (6) shows the results for the second order SETAR. The *V23* test shows that for both the fixed and adaptive smoothing there are no dramatic deviations of the rejection frequencies from the nominal level. The values of c seem not to play a relevant role in modeling the nonlinearity in the time series. As expected, the spurious ARCH effect discussed earlier is more serious in the higher order case. For the fixed bandwidth and $c = 0.5$ the rejections are 60% for T = 300 and 44% for T = 900. When c assumes the largest value the frequencies are 14% and 12% as the sample size increases. The benefit deriving from the adaptive bandwidth is now clear: for T = 300 the *HET* rejects in 36%, 17% and 8.7 for c that increases from the smallest to the largest value. Overall, the previous conclusion that c equal to 1 gives better results for the homoskedastic SETAR is confirmed also for the second order case. However, the adaptive bandwidth is now a crucial element to achieve reasonable rejection frequencies. The *HET* and *V23* tests are around 8% rejections for T = 300 and decrease toward the nominal level for larger samples.

Table (6) here

Table (7) shows the rejection frequencies for the SETAR model with dependence both in the first and second lag. The *SC* and *V23* tests have almost unit power to detect the linear and nonlinear dependence in the simulated model. The power against the *HET* alternative is somewhat higher than expected but probably the result of the misspefication of the forecasting densities. The linear parametric bootstrap accounts for most of the linear dependence in the data but the *V23* still points to the deficiency in modeling the nonlinearity. The previous discussion of the performance of the MFD largely holds also when the dependence occurs both in the first and second lag. The adaptive smoothing is important to correctly model the tails of the distribution and minimize the spurious ARCH effect.

Table (7) here

4.5 Discussion

The Monte Carlo results show that the Markov Forecasting Density (MFD) is a useful approach to model dependence in the data. It is able to capture both linear and nonlinear structures as the previous simulation exercise indicates. This is particularly helpful in a forecasting framework in which the main goal is to generate reliable predictions rather than understanding the underlying dynamics. The simulations also suggest some practical direction for the implementation of the method. First, the use of an adaptive bandwidth is a convenient method to deal with the "curse of dimensionality" in higher orders. It is advisable to use an adaptive smoothing scheme to implement the MFD. Another fact that emerge from the simulations is that excessive undersmoothing could lead to forecasting densities that are too concentrated. Although this helps in tracking closely the underlying evolution of the time series, it underestimates the variance of the forecasting density and shows up as periods of small values of the PIT followed by periods of large values. A safe choice for the bandwidth appears to be in the range 0.75 and 1 time the values of the adaptive bandwidth $h_T(adaptive)$. The method has reasonable performances already for in-sample length of 300, even though for some models a larger sample is required to have rejection frequencies close to the 5% level.

5 Empirical Application

As discussed in the Introduction, a successful field of application of nonlinear time series models has been to modeling business cycles. An early reference is the markov-switching model of Hamilton (1989) for the growth rate of real GDP. However, more recent work shows that the in-sample success of nonlinear models does not translate in improved out-of-sample predictability. Evaluating forecasting ability using the RMSPE (Root Mean Square Prediction Error), nonlinear models perform as well as linear models. The alternative discussed earlier is to evaluate the density forecasts rather than limiting the comparison to point forecasts. However, further problems arise as we argued in Section (3). Testing the uniformity of the PIT might have negligible power in detecting the dynamic misspecification of the forecasting densities. In this Section, we re-evaluate these results by using an extended battery of tests to detect misspecification of the linear (and nonlinear) model.

We consider the real GDP for U.S. from 1955 to the end of 2004. The series is at the quarterly frequency and relatively short for the nonparametric technique and the evaluation tests to give powerful answers. Hence, we consider also other variables that are interpreted as indicators of the business cycle condition: the Coincident Indicator (from the Conference Board) and Industrial Production. Both series are at the monthly frequency and the sample size increases to 532 observations. For all series

we consider the growth rate. Details of the sample period and the number of observations is given in Table (8). We chose to approximately split the full sample in half for estimation and the second half for forecasting. We forecast one-step ahead with the in-sample set expanding to include the new observations. For the MFD we select the order based on the $\delta(p_{MFD})$ test proposed in Diks and Manzan (2002) for the first available in-sample period. The results strongly indicate that 3 is the optimal order for the series considered. Based on the simulation results, we use a bandwidth equal to 0.75 the fixed and adaptive bandwidth values described in Section (2.3). To make our results comparable we also present the results for a linear AR specification (indicated *LIN*) and a two-regimes $SETAR(p_{TAR}, d)$ model defined as

$$Y_{t+1} = [X_t\theta_1 + \sigma_1\epsilon_{t+1}]I\left\{\left(\sum_{i=0}^{d-1} Y_{t-i}\right) \leq r\right\} + [X_t\theta_2 + \sigma_2\epsilon_{t+1}]I\left\{\left(\sum_{i=0}^{d-1} Y_{t-i}\right) > r\right\} \tag{9}$$

where Y_{t+1} denotes the growth rate of the series and X_t is a p_{TAR}-dimensional vector of lagged values of Y_{t+1}. The switching in the model depends on the cumulative growth rate in the last d months. The vectors θ_1 and θ_2 represent the parameters governing the dynamics in the two regimes. We allow for heteroskedastic regimes and denote the variances of the innovations by σ_1 and σ_2, respectively. We followed the approach of Siliverstovs and van Dijk (2003) and select recursively the lags p_{AR}, p_{TAR} and d based on a search up to lag 5 and the AIC criterion. The forecasting densities are then obtained by drawing with replacement from the standardized residuals of the models. The density evaluation is based on the tests introduced earlier with a lag order of 5 (for SC, HET, and $V23$).

Table (8) shows the p-values of the evaluation tests of the forecasting densities for the different series. We forecast the real US GDP growth rate series from the first quarter of 1980 until the forth quarter of 2004 (100 observations both in- and out-of-sample). The linear AR model performs poorly: the KS, HET and V23 reject the respective null hypothesis at 5% level while the SC test has a p-value of 0.06. This suggests that the LIN model does not account properly for the dynamics of the series. It neglects to account for nonlinear and heteroskedastic effects in the data. The SETAR and MFD perform significantly better with p-values larger than the significance level. These results suggest that the nonlinearity in the GDP growth is properly characterized by the regime-type dynamics of the SETAR model.

More discriminatory power can be achieved by considering monthly observations of economic variables often considered indicators of the business cycle situation. The variables mentioned above are for the period from January 1960 until April 2004 and we forecast out-of-sample starting from January 1986 (??? observations). The US Coincident Indicator is an index produced by the Conference Board

and it represents a weighted average of the information contained in four business cycle variables (employment, income, industrial production and sales). The target of this indicator is to measure the status of the cycle. Also for this variable, the linear forecasting densities shows signs of misspecification. The KS test rejects the null of uniformity of the PIT while the $V23$ has a p-value equal to 0.09. For this variable, the finding of neglected nonlinearity are less strong compared to the results for GDP growth. However, both SETAR and MFD seem to account more properly for the nonlinearity in the data. For the SETAR density forecast the KS test suggests the departure from uniformity of the distribution of the PIT. Instead, the MFD shows no signs of misspecification in the tests considered.
The third business condition indicator that we consider is Industrial Production. Table (8) indicates that both the KS and the $V23$ tests reject their null hypothesis for the AR density forecasts. Contrary to the results above, the SETAR density seem to leave some unexplained nonlinear features in the growth rate of IP. The $V23$ test for neglected nonlinearity rejects the null with a p-value equal to 0.03. A possible explanation for this rejection is the assumption of a two-regimes dynamics for the threshold model. In this case it is clear the advantage of considering a nonparametric approach: the MFD (in particular with adaptive bandwidth) does not indicate signs of misspecification in any of the tests considered.
This empirical application makes clear the two main contribution of the paper. First, considering nonparametric density forecasts is very useful in properly accounting for the dynamics in economic variables. On the one hand, linear models are proved to have poor ability to forecast the variables considered. On the other hand, nonlinear models require the specification of a functional form that might be inappropriate for the economic variable considered. In the regime-switching setup some choices to make are the number of regimes and the variable that governs the transition between regimes. As we argued earlier, in a forecasting setup the interpretability of the model might be less interesting compared to having a very flexible specification able to capture the many different nonlinear features of economic variables. The second contribution of the paper is to stress the importance of testing for the misspecification of the PIT. If we suspect that the dynamics of the variables is nonlinear, we need to evaluate the ability of the density forecasts to take that into account. In the application above, our findings would be very different if we disregard the $V23$ test. We would conclude that the AR forecasts show some departure from uniformity but account correctly for the dynamics of the US Industrial Production growth rate. We would also conclude that the SETAR is a proper forecasting model. However, the $V23$ test suggests that both AR and SETAR density forecasts miss to account for some nonlinear features in the dynamics of the business cycle variables. Thus, a valid alternative to consider is the nonparametric approach proposed in this paper.

6 Conclusions and Future Work

In this paper we propose a nonparametric bootstrap technique to predict the density of markovian time series. In addition, we also extend the testing approach for forecasting densities to include test against specific alternatives, such as autocorrelation, ARCH and linearity. By simulation we show that the method is successful in accounting for linear and nonlinear dependence. This is a useful property of the method, in particular in cases in which the analyst suspects the existence of nonlinearities in the data under investigation. In addition, the empirical application to business cycles variables suggests that the nonparametric method performs well for all them while parametric specification might fail to correctly identify the nonlinearity in the data.

However, the approach is quite new and requires more work on some relevant issues. In this paper we only consider a forecasting horizon of 1-step ahead. It would be relevant to examine the performance of the MFD on multi-step ahead forecasts. In this case, the comparison between direct and iterative schemes might show a difference in performance among them. Another interesting development is to investigate the performance of data-driven methods for bandwidth selection. The Monte Carlo experiments suggest that different models might require different degrees of smoothing. A data-driven selection of the degree of smoothing would automatically adapt to the need of the time series at hand.

In this paper we also proposed a testing framework based on the PIT by extending the use of standard tests used in econometrics. The testing strategy has the advantage of suggesting to the analyst the directions in which the forecasting model could be improved. More work should focus on elaborating tests that have high power also in small samples and for multistep horizons. Finally, the MFD method relies on nonparametric estimation of multivariate densities. As discussed earlier, when the markov order is high it requires large sample sizes to give reliable results. An improvement would be to impose some structure on the forecasting density such that the nonparametric problem is reduced to a lower dimension.

References

Cheng, B. and Tong, H. (1992). On consistent nonparametric order determination and chaos. *Journal of the Royal Statistical Society, Series* B, **54,** 427–449.

Clements, M.P., Franses, P.H., Smith, J. and van Dijk, D. (2003). On setar non-linearity and forecasting. *Journal of Forecasting*, **22,** 359–375.

Clements, M.P., Franses, P.H. and Swanson, N.R. (2004). Forecasting economic and financial time-series with non-linear models. *International Journal of Forecasting*, **20,** 169–183.

Clements, M.P. and Smith, J. (1997). The performance of alternative forecasting methods for setar models. *International Journal of Forecasting*, **13,** 463–475.

Clements, M.P. and Smith, J. (2000). Evaluating the forecast densities of linear and non-linear models: Application to output growth and unemployment. *Journal of Forecasting*, **19,** 255–276.

Corradi, V. and Swanson, N.R. (2004). Predictive density evaluation. In *Handbook of Economic Forecasting* (eds G. Elliott, C.W.J. Granger and A. Timmermann). forthcoming.

de Gooijer, J. and Zerom, D. (2003). On conditional density estimation. *Statistica Neerlandica*, **57,** 159–176.

de Gooijer, J.G. and Kumar, K. (1992). Some recent developments in non-linear time series modelling, testing and forecasting. *International Journal of Forecasting*, **8,** 135–156.

Diebold, F.X., Gunther, T.A. and Tay, A.S. (1998). Evaluating density forecasts. *International Economic Review*, **39,** 863–883.

Diks, C. and Manzan, S. (2002). Tests for serial independence and linearity using the correlation integrals. *Studies in Nonlinear Dynamics and Econometrics*, **6,** number 2.

Fan, J., Yao, Q. and Tong, H. (1996). Estimation of conditional densities and sensitivity measures in nonlinear dynamical systems. *Biometrika*, **83,** 189–206.

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

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

Granger, C.W.J. and Terasvirta, T. (1993). *Modelling Nonlinear Economic Relationships*. Oxford University Press.

Hamilton, J. (1989). A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica*, **57,** 357–384.

Hong, Y. and Li, H. (2005). Nonparametric specification testing for continuous-time models with applications to term structure of interest rates. *Review of Financial Studies*, **18,** 37–84.

Hong, Y., Li, H. and Zhao, F. (2004). Out-of-sample performance of discrete-time spot interest rate models. *Journal of Business and Economic Statistics*, **22,** 457–473.

Horowitz, J.L. (2003). Bootstrap methods for markov processes. *Econometrica*, **71,** 1049–1082.

Hyndman, R.J., Bashtannyk, D.M. and Grunwald, G.W. (1996). Estimating and visualizing conditional densities. *Journal of Computational and Graphical Statistics*, **5,** 315–336.

Hyndman, R.J. and Yao, Q. (2002). Nonparametric estimation and symmetry tests for conditional density functions. *Journal of Nonparametric Statistics*, **14,** 259–278.

Krolzig, H.M. (1997). *Markov-Switching Vector Autoregressions. Modelling, Statistical Inference and Application to Business Cycle Analysis*. Springer.

Paparoditis, E. and Politis, D.N. (2001). A markovian local resampling scheme for nonparametric estimators in time series analysis. *Econometric Theory*, **17,** 540–566.

Paparoditis, E. and Politis, D.N. (2002). The local bootstrap for markov processes. *Journal of Statistical Planning and Inference*, **108,** 301–328.

Pesaran, M.H. and Potter, S.M. (1997). A floor and ceiling model of US output. *Journal of Economic Dynamics and Control*, **21,** 661–695.

Rajarshi, M.B. (1990). Bootstrap in markov sequences based on estimates of transition density. *Annals of the Institute of Statistical Mathematics*, **40,** 565–586.

Siliverstovs, B. and van Dijk, D. (2003). Forecasting industrial production with linear, nonlinear, and structural change models. *Econometric Institute Report 2003-16*.

Terasvirta, T., Lin, C-F. and Granger, C.W.J. (1993). Power of the neural network linearity test. *Journal of Time Series Analysis*, **14,** 209–220.

Tong, H. (1990). *Non-linear Time Series*. Clarendon Press.

Table 1: **Power of the Predictive Density Evaluation Tests**

Model	*KS*		*SC*		*HET*		*V23*	
	P=100	P=300	P=100	P=300	P=100	P=300	P=100	P=300
AR(1) - ϕ=0.5	20	22	99	100	5.9	7.2	5.2	5.0
ARCH(1) - α=0.3	4.4	5.2	6.3	7.6	32	83	5.9	5.8
SETAR - hom. reg.	10	11	60	98	4.3	6.7	60	97
SETAR - het. reg.	8.9	10	24	56	11	15	74	99

Percentage of rejections of the null hypotheses of the tests at the 5% significance level based on 2000 simulations. The forecasting density is obtained by resampling the observations of the process. For SC, HET and $V23$ tests we use one lag.

Table 2: **Simulated Model:** *AR(1) - ϕ = 0.5*

T	P	Test	*LIN*	MFD fixed			MFD adaptive		
				c = 0.5	c = 0.75	c = 1	c = 0.5	c = 0.75	c = 1
300	300	*KS*	4.3	5.3	5.0	5.1	5.1	4.8	5.5
		SC	5.6	5.8	6.5	9.3	6.4	7.0	9.8
		HET	5.2	14	8.0	6.9	11	5.8	5.9
		V23	5.2	5.5	5.1	6.0	5.2	5.0	5.9
600	300	*KS*	4.3	4.7	4.9	4.8	4.3	5.5	4.5
		SC	5.3	5.8	6.7	7.9	6.2	7.4	8.9
		HET	4.5	11	6.4	5.8	7.3	5.8	4.9
		V23	4.0	4.8	5.6	4.8	5.2	5.6	4.7
900	300	*KS*	3.9	4.3	4.8	4.9	3.9	4.9	4.3
		SC	5.3	5.2	6.0	6.2	5.7	6.6	7.1
		HET	5.4	8.6	7.1	5.2	6.8	5.5	4.2
		V23	5.1	5.2	5.5	5.1	5.0	5.4	5.3

Percentage of rejections of the null hypothesis of the tests based on 2000 simulations. The number of bootstrap is equal to 1000 for all methods. The lag in the MFD, LIN and the tests is set equal to 1. Significance level for all tests is 5%.

Table 3: **Simulated Model:** *ARCH(1) - α = 0.3*

T	P	Test	*LIN*	MFD fixed			MFD adaptive		
				c = 0.5	c = 0.75	c = 1	c = 0.5	c = 0.75	c = 1
300	300	*KS*	5.6	6.8	5.7	4.4	6.5	5.6	4.9
		SC	9.7	6.3	4.4	3.7	6.0	4.0	3.3
		HET	83	23	19	17	17	15	14
		V23	5.9	5.6	5.4	4.4	5.1	4.8	4.0
600	300	*KS*	4.7	4.2	4.6	5.1	4.3	4.3	4.9
		SC	8.0	5.1	4.7	5.8	4.4	4.5	5.6
		HET	84	18	15	12	13	12	11
		V23	6.5	5.5	5.8	4.3	5.5	5.2	4.4
900	300	*KS*	4.8	5.2	4.8	4.3	5.0	4.8	4.1
		SC	7.2	4.7	5.3	4.8	4.9	5.1	5.0
		HET	84	15	12	9.6	11	9.8	7.8
		V23	6.3	5.3	5.1	5.9	5.3	5.2	5.1

Percentage of rejections of the null hypothesis of the tests based on 2000 simulations. The number of bootstrap is equal to 1000 for all the methods. The lag in the MFD, LIN and the tests is set equal to 1. Significance level for all tests is 5%.

Table 4: **Simulated Model:** *SETAR(1) - homoskedastic regimes*

T	P	Test	*LIN*	MFD fixed			MFD adaptive		
				c = 0.5	c = 0.75	c = 1	c = 0.5	c = 0.75	c = 1
300	300	*KS*	7.0	4.3	4.2	3.9	4.8	4.7	4.3
		SC	12	4.9	4.9	3.7	4.7	4.9	4.0
		HET	8.4	17	12	7.6	16	11	7.0
		V23	95	4.9	6.3	6.3	4.7	6.7	8.1
600	300	*KS*	5.1	5.0	3.4	3.6	4.8	3.4	3.7
		SC	13	5.2	4.8	4.5	5.1	5.7	4.9
		HET	8.7	11	8.8	7.3	10	7.4	6.5
		V23	94	4.8	5.2	6.9	5.3	5.6	7.7
900	300	*KS*	6.9	4.9	4.5	4.2	5.2	3.9	4.5
		SC	13	5.1	4.7	5.2	5.1	4.5	5.6
		HET	8.2	9.1	6.9	6.5	8.4	6.8	6.0
		V23	94	4.9	4.7	6.5	4.9	4.5	6.9

Percentage of rejections of the null hypothesis of the tests based on 2000 simulations. The number of bootstrap is equal to 1000 for all the methods. The lag in the MFD, LIN and the tests is set equal to 1. Significance level for all tests is 5%.

Table 5: **Simulated Model:** *SETAR(1) - heteroskedastic regimes*

T	P	Test	*LIN*	MFD fixed			MFD adaptive		
				c = 0.5	c = 0.75	c = 1	c = 0.5	c = 0.75	c = 1
300	300	*KS*	7.6	4.4	4.9	5.3	4.3	5.3	5.2
		SC	10	5.9	5.9	5.5	5.5	5.5	5.1
		HET	15	19	14	8.3	16	8.3	7.3
		V23	99	5.7	7.9	12	5.6	12	13
600	300	*KS*	7.2	4.3	5.0	4.6	4.1	5.1	5.6
		SC	11	5.8	5.6	7.0	5.9	5.1	6.7
		HET	16	14	9.6	6.9	12	8.8	6.0
		V23	100	6.0	7.0	9.1	6.3	7.5	10
900	300	*KS*	8.3	4.8	4.2	5.0	4.4	4.0	5.0
		SC	8.7	5.9	5.8	6.0	5.8	5.7	5.3
		HET	17	12	8.6	7.2	11	8.1	6.8
		V23	98	6.3	6.1	8.5	6.1	6.7	10

Percentage of rejections of the null hypothesis of the tests based on 2000 simulations. The number of bootstrap is equal to 1000 for all the methods. The lag in the MFD, LIN and the tests is set equal to 1. Significance level for all tests is 5%.

Table 6: **Simulated Model:** *SETAR(2) - homoskedastic regimes*

T	P	Test	IND	*LIN*	MFD fixed			MFD adaptive		
					c = 0.5	c = 0.75	c = 1	c = 0.5	c = 0.75	c = 1
300	300	*KS*	12	6.7	54	10	5.5	46	8.4	5.3
		SC	96	11	5.9	6.0	5.2	5.8	5.7	4.5
		HET	6.2	7.1	60	31	14	36	17	8.7
		V23	85	78	6.4	5.7	7.5	6.7	5.3	8.0
600	300	*KS*	12	6.1	27	6.9	5.2	21	5.9	5.4
		SC	96	9.6	6.3	5.0	4.8	5.8	5.1	4.3
		HET	5.8	6.7	50	24	12	28	13	7.9
		V23	85	80	7.6	6.6	8.3	6.0	5.9	8.5
900	300	*KS*	12	6.3	16	5.5	5.3	13	5.1	5.1
		SC	96	8.9	5.8	5.5	4.4	5.6	5.2	4.4
		HET	6.3	7.0	44	20	12	24	12	7.2
		V23	86	80	7.1	7.1	7.0	6.4	6.8	6.6

Percentage of rejections of the null hypothesis of the tests based on 2000 simulations. The number of bootstrap is equal to 1000 for all the methods. The lag in the MFD, LIN and the tests is set equal to 2. Significance level for all tests is 5%.

Table 7: **Simulated Model:** *SETAR(1-2) - homoskedastic regimes*

T	P	Test	IND	*LIN*	MFD fixed			MFD adaptive		
					c = 0.5	c = 0.75	c = 1	c = 0.5	c = 0.75	c = 1
300	300	*KS*	3.3	5.9	34	5.2	3.9	29	5.3	4.6
		SC	100	11	6.5	8.2	12	7.0	8.7	14
		HET	23	28	47	21	11	29	12	7.1
		V23	99	93	6.3	6.9	12	6.4	7.4	14
600	300	*KS*	3.7	5.4	15	5.0	4.4	14	4.5	5.3
		SC	100	13	6.9	7.6	9.8	6.3	8.5	11
		HET	23	28	41	16	9.2	22	9.4	6.1
		V23	97	92	7.9	8.3	9.2	6.9	8.5	11
900	300	*KS*	3.4	5.7	11	3.9	3.3	8.5	4.0	3.6
		SC	100	13	6.6	7.0	10	7.3	7.7	11
		HET	23	27	32	13	8.7	17	7.2	6.4
		V23	98	93	7.3	7.9	9.8	7.4	7.3	11

Percentage of rejections of the null hypothesis of the tests based on 2000 simulations. The number of bootstrap is equal to 1000 for all the methods. The lag in the MFD, LIN and the tests is set equal to 2. Significance level for all tests is 5%.

Table 8: Comparison of Density Forecasting Models

Series	T	P	Order	Method	KS	SC	HET	V23
US real GDP	1955(1)/1979(4)	1980(1)/2004(4)	3	LIN	0.03	0.06	0.01	0.01
(quarterly)	T = 100	P=100		SETAR	0.26	0.07	0.14	0.27
				MFD (fixed)	0.35	0.58	0.15	0.16
				MFD (adapt)	0.18	0.41	0.30	0.58
US Coincident Indicator	1960(1)/1985(12)	1986(1)/2004(4)	3	LIN	0.01	0.30	0.18	0.09
(monthly)	T = 312	P=220		SETAR	0.01	0.12	0.10	0.28
				MFD (fixed)	0.14	0.30	0.67	0.12
				MFD (adapt)	0.14	0.44	0.58	0.35
US Industrial Production	1960(1)/1985(12)	1986(1)/2004(4)	3	LIN	0.02	0.61	0.34	0.04
(monthly)	T = 312	P=220		SETAR	0.18	0.53	0.32	0.03
				MFD (fixed)	0.91	0.10	0.24	0.06
				MFD (adapt)	0.98	0.17	0.18	0.21

p-values of the density forecasts tests described in Section (3). For all tests we used a lag order of 5. The bandwidth for the MFD is fixed for all series to 0.75 the values of the rule-of-thumb and adaptive given in Section (2.3). The markov order selected is indicated in the column *order*. The lag orders of the LIN and SETAR forecasting methods are chosen performing a search up to lag 5 and by the AIC selection criterion.